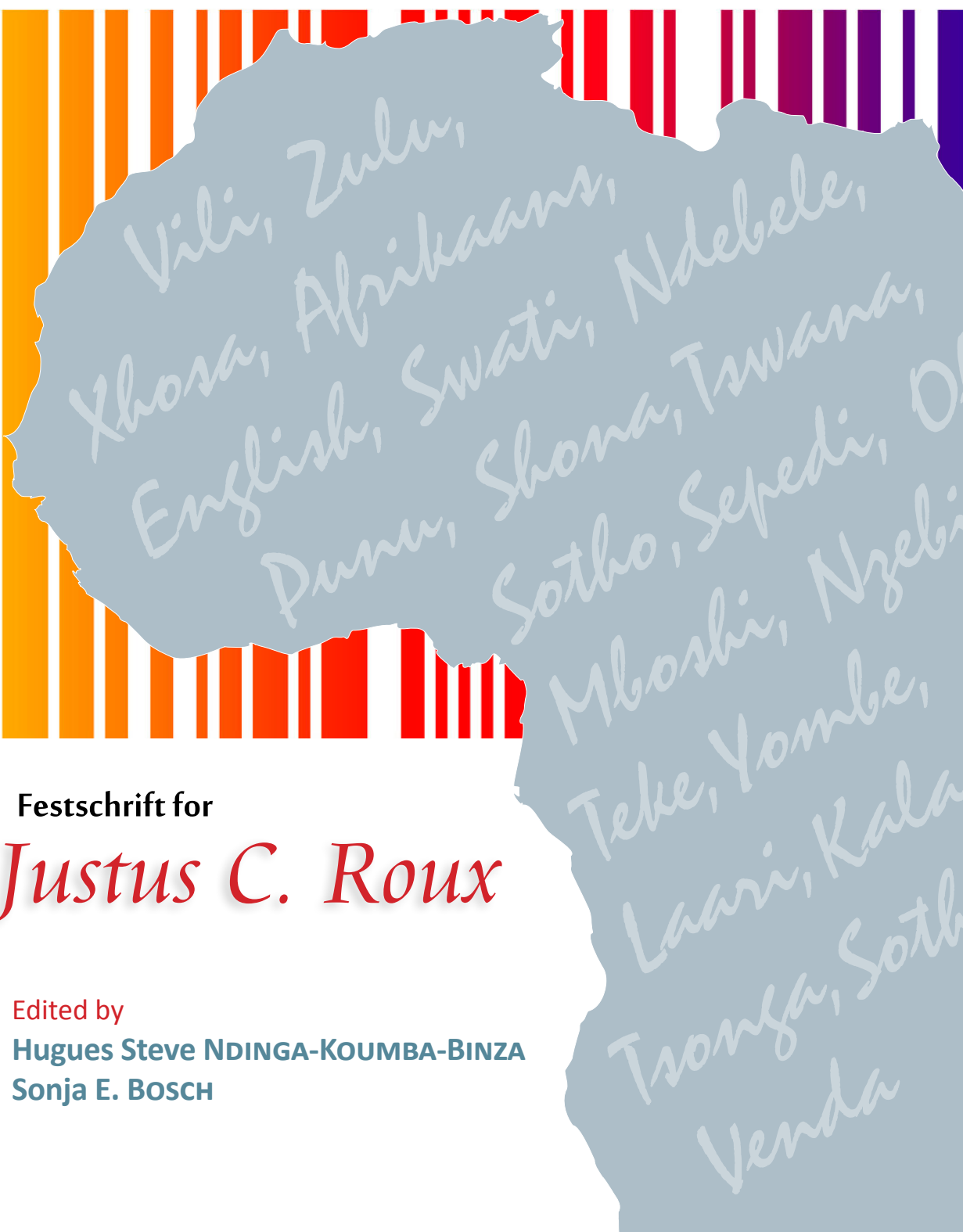


# Language Science and Language Technology in Africa



Vili, Zulu,  
Xhosa, Afrikaans,  
English, Swati, Ndebele,  
Duma, Shona, Tswana,  
Sotho, Sepedi, O  
Mboshi, Nzebi  
Teke, Yombe,  
Lauri, Kala  
Tsonga, Soth  
Venda

Festschrift for

*Justus C. Roux*

Edited by

Hugues Steve NDINGA-KOUMBA-BINZA

Sonja E. BOSCH



# Language Science and Language Technology in Africa:

A Festschrift for Justus C. Roux

Edited by

Hugues Steve NDINGA-KOUMBA-BINZA

&

Sonja E. BOSCH



*Language Science and Language Technology in Africa: Festschrift for Justus C. Roux*

Copyright © 2012 SUN MeDIA Stellenbosch and contributing authors

All rights reserved.

No part of this book may be reproduced or transmitted in any form or by any electronic, photographic or mechanical means, including photocopying and recording on record, tape or laser disk, on microfilm, via the Internet, by e-mail, or by any other information storage and retrieval system, without prior written permission by the publisher.

First edition 2012

ISBN: 978-1-920338-79-4

e-ISBN: 978-1-920338-80-0

DOI: 10.18820/9781920338800

Set in 10/12 Palatino Linotype

Cover design: Liezel Meintjes

SUN PRESS is an imprint of SUN MeDIA Stellenbosch. Academic, professional and reference works are published under this imprint in print and electronic format. This publication may be ordered directly from [www.sun-e-shop.co.za](http://www.sun-e-shop.co.za).

Printed and bound by SUN MeDIA Stellenbosch, Ryneveld Street, Stellenbosch, 7600.

[www.africansunmedia.co.za](http://www.africansunmedia.co.za)

[www.sun-e-shop.co.za](http://www.sun-e-shop.co.za)

## ACKNOWLEDGEMENT TO REVIEWERS

Every contribution in this volume was subjected to a double blind peer-review process. We wish to acknowledge the participation of the following colleagues in the production of this book. Some of them had to review the same chapter two or three times, and others reviewed more than one chapter. Thank you!

Mariëtta Alberts (NWU)	Shamila Naidoo (UKZN)
Etienne Barnard (NWU)	Steve Ndinga-Koumba-Binza (NWU)
Ian Bekker (NWU)	Thomas Niesler (SU)
Ansu Berg (NWU)	Mildred Nkolola-Wakumelo (Zambia)
Sonja E. Bosch (UNISA)	Dion Nkomo (RU)
Willem Botha (WAT)	Blanche Nyangone Assam (UWC)
Emmanuel Chabata (Zimbabwe)	Sulene Pilon (NWU)
Febe de Wet (CSIR)	Rigardt Pretorius (NWU)
Johan du Plessis (WAT)	Danie Prinsloo (UP)
Rufus Gouws (SU)	Martin Puttkammer (NWU)
Marissa Griesel (NWU)	Aditi Sharma Grover (CSIR)
Hendrik Groenewald (NWU)	Frenette Southwood (SU)
Maxwell Kadenge (WITS)	Elsabé Taljard (UP)
Inge Kosch (UNISA)	Charl van Heerden (CSIR)
Albert E. Kotzé (UNISA)	Gerhard van Huyssteen (NWU)
John Lubinda (Botswana)	Linda van Huyssteen (TUT)
Ludwine Mabika Mbokou (Gabon)	Bertus van Rooy (NWU)
Paul Achille Mavoungou (Gabon)	Daan P. Wissing (NWU)
Cindy A. McKellar (NWU)	Sabine Zerbian (Germany)

The editors are also grateful to Professor Axel Fleisch (Department of World Cultures, University of Helsinki, Finland) and to Dr Johan C.M.D. du Plessis (former editor of *Lexikos*, Bureau of the Woordeboek van die Afrikaanse Taal, South Africa) for their thorough and insightful review of the entire volume.

## CONTENTS

JUSTUS C. ROUX'S BRIEF CURRICULUM VITAE	v
<b>PREFACE</b>	xi
<b>PART 1: PHONETICS AND PHONOLOGY</b>	
<b>STRESS ASSIGNMENT IN BLACK SOUTH AFRICAN ENGLISH</b>	
Sabine Zerbian	1
<b>A COMPARATIVE STUDY OF EXPLOSIVE [b] AND IMPLOSIVE [ɓ] IN NGUNI</b>	
Shamila Naidoo	21
<b>ON THE PHONETIC AND THE PHONOLOGICAL STUDIES OF SOUTHERN AFRICAN LANGUAGES</b>	
John Lubinda	33
<b>ON THE MANY FACES OF TONE IN SOUTHERN SOTHO: A CASE STUDY</b>	
Daan Wissing	47
<b>SYLLABIFICATION OF CONSONANTS IN SESOTHO AND SETSWANA</b>	
Mildred Nkolola-Wakumelo, Liketso Rantso, Keneilwe Matlhaku	61
<b>PHONETIC DATA AND PHONOLOGICAL THEORY: A REPORT FROM THE CIVILI VOWEL DURATION ISSUE</b>	
Hugues Steve Ndinga-Koumba-Binza	83
<b>PART 2: LANGUAGE DESCRIPTION AND RESOURCES</b>	
<b>THE ORATURE-GRAMMAR INTERFACE: ON "RHYMES" IN AFRICAN ORAL VERBAL ART</b>	
H. Ekkehard Wolff	99
<b>BOOTSTRAPPING THE DEVELOPMENT OF MORPHOLOGICAL ANALYSERS FOR "DISPERSED" NGUNI LANGUAGES - A LINGUISTIC INVESTIGATION</b>	
Sonja E. Bosch	123
<b>FROM 'BELEAGUERED' TO 'EMANCIPATED' ORTHOGRAPHY: THE CASE OF NORTHERN SOTHO</b>	
Inge Kosch	145
<b>A SURVEY OF BILINGUALISM IN MULTILINGUAL GABON</b>	
Ludwine Mabika Mbokou	163

**PART 3: LEXICOGRAPHY AND TERMINOLOGY**

<b>TOWARDS A YILUMBU DICTIONARY OF IDIOMATIC PHRASES</b>	
Paul Achille Mavoungou	177
<b>LANGUAGE STANDARDISATION AND OTHER LANGUAGE AND CONTENT RESOURCES: SABS TC 37 - A MIRROR COMMITTEE OF ISO/TC 37</b>	
Mariëtta Alberts	199
<b>ENRICHING A DICTIONARY DATABASE WITH MULTI WORD EXPRESSIONS</b>	
Thapelo Otlogetswe	215
<b>DICTIONARY BASIS AND LEMMATISATION FOR AN ENCYCLOPEDIA DICTIONARY OF YILUMBU</b>	
Hugues Steve Ndinga-Koumba-Binza and Gilles Saphou-Bivigat	237
<b>INNOVATIVE STRATEGIES IN MACROSTRUCTURAL CHOICES</b>	
Rufus H. Gouws	251

**PART 4: HLT RESEARCH AND DEVELOPMENT**

<b>FREQUENCY-BASED DATA SELECTION FOR STATISTICAL MACHINE TRANSLATION WITH SCARCE RESOURCES</b>	
Cindy A. McKellar and Hendrik J. Groenewald	271
<b>VOICE USER INTERFACE DESIGN FOR EMERGING MULTILINGUAL MARKETS</b>	
Gerhard van Huyssteen, Aditi Sharma Grover, Karen Calteaux	291
<b>THE RELATIONSHIP BETWEEN THE AUTOMATIC ASSESSMENT OF ORAL PROFICIENCY AND OTHER INDICATORS OF FIRST YEAR STUDENTS' LINGUISTIC ABILITIES</b>	
Febe de Wet, Thomas Niesler, Christa van der Walt	309
<b>CONCEPTS FOR DIFFERENT TYPES OF INFORMATION TOOLS</b>	
Henning J. Bergenholtz	327
<b>'MARKETSPEAK' IN IGBO: A SPEECH SYNTHESIS TRAINING PROJECT</b>	
Dafydd Gibbon, Ugonna Duruibe, Jolanta Bachan	339
<b>INDEX</b>	360



### **JUSTUS C. ROUX'S BRIEF CURRICULUM VITAE**

- 1947** Born 15 January 1947 in Brakpan, South Africa.
- 1967** Received BA degree majoring in Afrikaans and Dutch, Northern Sotho and Zulu from the then Potchefstroom University for Christian Higher Education (PUCHE).
- 1968** Received an Honours degree in African languages from PUCHE.
- 1969** Appointed Junior Lecturer in African languages at PUCHE.
- 1971** Received a Master of Arts in African languages from PUCHE.
- 1971** Promoted to Lecturer in African languages at PUCHE.
- 1972** Appointed Lecturer in African languages at Stellenbosch University. This same year, he married Cornelia and they have three grown-up children – Ruzanne, Christelle and Justus – and two grandchildren.
- 1979** Received his Doctor Litteratum (DLitt) degree in African Languages and General Linguistics from Stellenbosch University.



- 1980** Promoted to Senior Lecturer in African languages at Stellenbosch University. At the same period he became member of the Suid-Afrikaanse Akademie vir Wetenskap en Kuns; a membership that he still holds to date. He also took on a membership of the African Language Association of Southern Africa (ALASA), an organisation of which he is still a member and has held various executive positions ranging from Chairperson of the Western Cape Chapter (1980, 1984-1988), Adjudicator for ALASA prizes for various publishers (1986), Member of the Board and Executive of ALASA (1986-1989, 1995-1997), and Secretary to the National Executive Board of ALASA (1995-1997).
- 1981-2004** He was member of the Linguistics Society of Southern Africa (LSSA).
- 1982** Appointed Head of the Department of African Languages at Stellenbosch. He was later re-appointed to this position five times until 2000.
- 1983** Appointed Member of the Faculty Executive Committee, Faculty of Arts at Stellenbosch University. He had this position renewed three times until 1999.
- 1983-1985** Member of Faculty Committee on Student Feedback.
- 1984** Attended two specialised international training courses. The first course, on "Theory and practice of Suggestopedic teaching", was held at the Institute for Language Teaching of the University of Stellenbosch and presented by Prof Charles Schmidt of Lind Institute, San Francisco, USA. The second course, on "Methods and Techniques of Accelerative Learning", was held by the Department of Continuing Education at the University of Houston, Houston, USA.
- 1985** Promoted to Full Professor (Ad Hominem) in African Languages at Stellenbosch University. During the same year, he also attended an international training course held by the Institute for Language Teaching of the University of Stellenbosch on "Educational Cognitive Styles - the development of creative thinking skills", presented by Prof J. Hand (USA) and Ms Beatrice Capdeville (Venezuela).
- 1986-1990** Appointed member of the Management Committee of the Institute for Language Teaching at Stellenbosch University. He chaired the Committee from 1989 to 1990.

- 1988-1990** Appointed member of the Senate Sub-Committee for Language Laboratories. It is also in this period that he was appointed member of the Selection Committee of the Human Sciences Research Council on Modern Languages, Arts and Information Science, responsible for adjudication of national research and travel grants from 1989 to 1990. He was simultaneously (until 1991) member of the Management Committee of the National Working Group on Computational Linguistics, functioning under the auspices of the Human Sciences Research Council.
- 1990** Elected Chairperson of the Committee for Language Laboratories. He was re-elected in 1991 and in 1992.
- 1993-2004** Founded an NRF unit and became its Director, i.e. the Research Unit for Experimental Phonology at the University of Stellenbosch (RUEPUS). It is during this period that he became Chairperson of the Special Interest Group on Laboratory Phonology functioning under the auspices of the African Language Association of Southern Africa. He was simultaneously founder member of the South African Foundation for Language and Speech Technology Development in 1999 until the organisation discontinued its activities in 2004.
- 1994** Became member of the Sub-Committee A (Arts & Social Sciences) of the Research Committee of Stellenbosch University until 1999.
- 1994-1995** Was member of the Management Committee of the Bureau for Continuing Education at Stellenbosch University.
- 1996** In November 1996, he worked at the Institute for Computational Linguistics in Stuttgart, Germany.
- 1997-1998** Elected Chairperson of the Management Committee for Computer Based Education in the Humanities. The following year he became Chairperson of the Management Committee of the Computer Use in the Humanities (HUMARGA) until 2003.
- 1998-1999** Appointed Vice-Dean of the Faculty of Arts at Stellenbosch University.
- 1999** Founding member of the Special Interest Group for Language and Speech Technology of the African Language Association of Southern Africa (ALASA-SIG).
- 2000** Appointed member of the Steering Committee for Language and Information Technology Development by the Department of Arts,

Culture, Science and Technology (DACST) and the Pan South African Language Board (PanSALB).

- 2001** Was once again Vice-Dean of the Faculty of Arts at Stellenbosch University from September to December 2001. This same year he acquired life membership of the Academy of Science of South Africa. It is also in this year that he was appointed Co-ordinator of the Advisory Panel on Human Language Technology Development in South Africa by the Minister of Arts, Culture, Science and Technology; he served in this position until 2002.
- 2002** Became member of the Academic Research Rating Panel for Language and Linguistics of the National Research Foundation.
- 2002 – 2007** Appointed Professor of African Languages at Stellenbosch University on a part-time contract.
- 2003 – 2006** Appointed Chairperson of the South African Technical Committee for the Standardisation of Terminology and other resources SABS TC/37. It is in this capacity that he was member of the South African delegation to the Annual International Meetings of ISO TC/37 in 2005 in Warsaw, Poland, in 2006 in Beijing, China, in 2007 in Provo, USA, and in 2010 in Dublin, Ireland.
- 2003-2009** Appointed member of the Steering Committee for Human Language Technologies Implementation by the Minister of Arts, Culture, Science and Technology.
- 2005-2008** Appointed Director of the newly founded Stellenbosch University Centre for Language and Speech Technology (SU-CLaST).
- 2005** Served as part-time Director of CatchWord Language and Speech Technologies (Pty) Ltd (a Stellenbosch University spin-off company).
- 2009-2010** Appointed part-time Senior Researcher at SU-CLaST. In the same period, he was nominated member of the HLT Expert Panel (HLTEP) by the Minister of the Department of Arts and Culture (DAC) for a period of five years to assist DAC in deploying HLT Research and Development projects in South Africa. In 2010, the Minister also appointed him Co-ordinator of a Task Team to develop a Blueprint for a Resource Management Agency for the National Centre for Human Language Technologies (NCHLT-RMA).

- 2009-2011** Appointed Senior Researcher at the Centre for Text Technology (CTexT) on the Potchefstroom Campus of the North-West University. Since 2009 he has been member of the Committee for Advanced Degrees, Research Unit for Language and Literature within the South African Context, and member of the Executive Committee of the Research Unit for Language and Literature within the South African Context since 2011.
- 2010** Since 2010, he is the Vice-President of the African HLT Association.
- 2011** Appointed Director of the Research Unit for Languages and Literatures in the South African Context on the Potchefstroom Campus of the North-West University as from May 2011. He is also member of the Advisory Panel for NRF Researcher Rating (Humanities and Social Sciences) since the same year.

## PREFACE

The genesis of this book dates from January 2010, when the editors met on the occasion of the first National HLT Network workshop at the Council of Scientific and Industrial Research (CSIR) in Pretoria, and agreed to compile a volume in honour of Professor Justus Roux to celebrate his 65<sup>th</sup> birthday in 2012. When suggested to a number of colleagues, the idea was warmly welcomed and many contributed, as is evident in the contents of the book.

The number of contributions (20 papers with 29 contributors), the calibre of the contributors (half a dozen are NRF-rated researchers, 10 internationally known professors and 4 world leaders in their respective fields), as well as the diversity of contributors' origins (9 African and European countries, i.e. Botswana, Denmark, Gabon, Germany, Lesotho, Nigeria, South Africa, Poland and Zambia) are testimony to the academic stature of the man being honoured through this book. Both his personality and his work have had an impact on the lives and careers of many of his friends and colleagues.

**Professor Justus Christiaan Roux** has, in fact, been the study leader, the promoter, the mentor, the colleague, or simply the friend of many worldwide. Close friends of his and colleagues are cited among the world top scholars. This explains the 100 plus visits he has made to top universities in Europe, America and Asia not only for world class conferences and workshops, but also for research stays, partnership and board meetings of several organisations such as the International Phonetics Association (IPA), the International Speech Communication Association (ISCA), the International Committee of Written Language Resources (WRITE) and the International Standards Organization (ISO) to name a few.

Born on 15 January 1947 in Brakpan (now part of the Ekurhuleni Metro Municipality), South Africa, Justus Roux attended the Carletonville High School from 1960 to 1964, before enrolling for a BA degree in the (then) Potchefstroom University for Christian Higher Education (now Potchefstroom Campus of the North-West University). He graduated in 1967 majoring in Afrikaans and Dutch, Northern Sotho and Zulu. He later obtained from the same institution an Honours degree (1968) and a Master of Arts in African Languages (1971). His research and scientific education has been marked with several training courses (e.g. Suggestopedic teaching, Accelerative learning, language technologies, etc.) throughout his long career that started as an academic in the position of Junior Lecturer in African Languages at Potchefstroom University for Christian Higher Education in 1969. He later became Lecturer in African Languages from 1969 to 1971 in that same institution.

In 1972 he moved to Stellenbosch University where he was successively Lecturer, Senior Lecturer and Professor of African Languages before retiring from teaching

in 2002. He was the Chair of the Department of African Languages at Stellenbosch six times (1982, 1986, 1988, 1992-1993, 1997-1999, and 2000). He was also the Vice Dean of the Faculty of Arts at Stellenbosch University for two different terms (1998-1999 and 2001-2002). In 1993, he founded the Research Unit for Experimental Phonology at the University of Stellenbosch (RUEPUS) which contributed in sparking laboratory phonology and language technologies research and development in South Africa. One of the major outcomes produced in RUEPUS was the successful development of a multilingual African Speech Technology system (i.e. a fully automated telephone-based multilingual query and booking system) for five official languages of South Africa, i.e. South African English, Zulu, Xhosa, Southern Sotho and Afrikaans.

Through the African Language Association of Southern Africa Special Interest Group (ALASA-SIG) on human language technologies that he co-established, he promoted human language technologies research and development not only in South African institutions (universities, private research organisations and companies) but also in other African countries such as Gabon, Ghana and Morocco together with a number of colleagues. It is in this trend that RUEPUS merged with the Digital Speech Processing Group (DSP Group) of the Stellenbosch University Department of Electrical and Electronic Engineering, with which it had already had an active collaboration over a period of twenty years, to form the Stellenbosch University Centre for Language and Speech Technology (SU-CLaST) in 2005. Justus Roux was subsequently appointed Director of SU-CLaST that same year.

His work and passion for promoting human language technologies in the African context led him to organise, co-organise or host conferences and workshops nationally and abroad in the field or related fields. One would note the following to mention the four most recent:

- (i) Special sessions and meetings on Human Language Technologies at the 13<sup>th</sup> International Conference of the African Language Association of Southern Africa (ALASA). University of Johannesburg, South Africa, 4–7 July 2005.
- (ii) the International Speech Communication Association's (ISCA) Tutorial and Research Workshop on Multilingual Speech and Language Processing (*MULTILING 2006*), Stellenbosch, 9-11 April 2006
- (iii) the pre-conference workshop on Networking the development of resources for African languages, at the 5th International Conference on Language Resources and Evaluation (LREC 2006), 21-28 May, 2006. Genoa, Italy.
- (iv) the HLT research and development workshop: "*Developing Human Language Technologies in Africa: a case for Gabon*", 26-28 November 2008. Institut de Recherche en Sciences Humaines (IRSH), Centre National

de Recherche Scientifique et Technologique (CENAREST), Libreville, Gabon.

Justus Roux's dedication to the development of language resources and language technologies in South Africa is reflected in his activities on a high level, e.g. his role of co-ordinator of the Advisory Panel on Human Language Technology Development in South Africa, from 2001 to 2002; and subsequently his membership of the Steering Committee for Human Language Technologies Implementation from 2003 – 2009. Both appointments were made by the Minister of Arts, Culture, Science and Technology. In 2009 he was nominated as member of HLT Expert Panel (HLTEP) by the Minister of the Department of Arts and Culture (DAC) for a period of five years to assist the DAC in deploying HLT Research and Development projects in South Africa. Based on his experience in the field of resource management and his collaboration with international language resource projects, he was also appointed co-ordinator of a Task Team to develop a Blueprint for a Resource Management Agency for the National Centre for Human Language Technologies (NCHLT-RMA) in 2010.

This volume brings together work from the fields of phonetics and phonology, morphology, sociolinguistics, terminology, lexicography and language technology research as a reflection of Justus Roux's wide range of academic interests. The common denominator in the majority of contributions is their reference to African languages (including Black South African English). The chapters in this book are organised in four topical sections. The first is concerned with *phonetics and phonology*; the second with various field studies grouped into the topic of *language description and resources*, i.e. morphology, sociolinguistics etc.; the third with *terminology and lexicography*; and the fourth with *language technology* research.

The first section contains six chapters. Chapter 1, by Sabine Zerbian, is a study of stress assignment in Black South African English (BSAE) at the word level. The study uses new empirical data to test the stress algorithm for BSAE proposed in previous studies such as van Rooy (2002) who claimed that there are indicators of a single stable system for stress assignment in Tswana English across speakers. Among other findings, the study confirms that the algorithm can account for many of the observed stress patterns and thereby refutes, in line with van Rooy (2002), the previously held impression that stress assignment in this variety is either idiosyncratic or restricted to the penultimate syllable.

Chapter 2, a contribution by Shamila Naidoo, is a comparative analysis of the explosive [b] and implosive [ɓ] consonants in three Nguni languages, i.e. Swati, Xhosa and Zulu. An experimental method and a corpus of minimal pairs are used to conduct a qualitative and quantitative examination of the parameters burst amplitude and closure duration in order to determine the degree of similarity between the studied consonantal sounds in the three languages.

Chapter 3 is a contribution by John Lubinda who protests against the weakness of phonetic and phonological descriptions of the majority of languages in Southern Africa. He pleads the case for a more scientific approach to the study of the sounds and sound systems of these languages, using available technical devices of the phonetics laboratory that guarantee a greater degree of objective precision, and applying appropriate phonological models that ensure descriptive adequacy.

Chapter 4, which is the only tone study of the volume, is a proposal by Daan Wissing. It is an acoustic description and an assessment of tonological characteristics of Southern Sotho. Findings strongly suggest that tone in Southern Sotho is no clear-cut phenomenon, as described or assumed in various writings.

Chapter 5 is a joint contribution by Mildred Nkolola-Wakumelo, Liketso Rantso and Keneilwe Matlhaku whose paper discusses the occurrence of derivative syllabic consonants in Sesotho and Setswana. The authors apply theoretical perspectives drawn from Distinctive Feature Phonology in order to show that the derivative syllabic consonant is characteristic of these two languages both belonging to the Sotho-Tswana group. This chapter is a collaboration of three colleagues based in three different Southern African countries, namely Botswana, Lesotho and Zambia.

Chapter 6, by Hugues Steve Ndinga-Koumba-Binza, concludes this first section with a report on the interaction between phonetic data and phonological theories with reference to the study conducted on the Civili vowel duration issue. The paper defines the nature of efficient data for a systematic study within the framework of the so-called *Phonetics-Phonology Interface Debate* (P-PID).

Section two comprises four chapters. In chapter 7, Ekkehard Wolff makes a study of the orature-grammar interface with reference to rhymes in African oral verbal art. In fact, the existence of “rhyming” as a salient aesthetic device has long been negated in traditional African verbal art, apart from assumed copies of Arabo-Islamic models. It is herein shown that, in the Chadic language *Lamang* in Nigeria, rhyme patterns are also salient in elevated discourse and narratives. A number of illustrative examples from two different speakers and different discourse genres are introduced and discussed.

Chapter 8, by Sonja Bosch, is a linguistic investigation of the feasibility of bootstrapping the development of morphological analysers for two ‘dispersed’ Bantu languages, by using an existing prototype of a Zulu morphological analyser. Cross-linguistic morphological similarities and distinguishing grammatical features between Zulu and two ‘dispersed’, resource-scarce Nguni languages, namely Zimbabwe Ndebele (S44) and Tanzanian Ngoni (N12) are examined with reference to their significance for bootstrapping purposes. The investigation focuses on the morphotactics and the morphophonological alternations of the languages involved.



In chapter 9, the contribution by Inge Kosch aims to illustrate how the orthography for Northern Sotho gradually freed itself from impractical and elaborate conventions to become a more user friendly writing system. The European missionaries who reduced the language to writing were informed by various factors in their choice of symbols, ranging from simple transfer from their source language on the one hand to scientifically motivated choices on the other. A high premium was placed on a practical orthography that was close to a phonetic orthography. This approach, however laudable, proved to be too technical for implementation by its target users.

In chapter 10, Ludwine Mabika Mbokou reflects on the notions of bilingualism in the multilingual context of Gabon, a French speaking state in central Africa. A commonly held view among the Gabonese people is that bilingualism only refers to a person who is able to speak two different European languages such as English and French. The author shows that bilingualism is a relative concept with specific reference to Gabonese children who, individually raised in a bilingual environment, are introduced to a second and even a third language at either a later or earlier stage of their childhood. As a result, the younger Gabonese generations are divided into two major groups where French and the native languages are wrestling for the initial language position.

Section three comprises five chapters. Chapter 11 by Paul Achille Mavoungou is an account of the planning of a Yilumbu idiomatic Dictionary. In order to discuss Yilumbu idiomaticity, the author restricts his discussion to the following points: (i) the methodology and theoretical assumptions of the work, and (ii) the type of oral traditions. A comparative terminology analysis of idiomatic expressions in English and Yilumbu is subsequently made.

In chapter 12, Mariëtta Alberts gives an overview of matters such as language development, standardisation of language, standard languages, harmonisation of languages, modernisation of languages and sociolinguistic factors regarding term creation. Emphasis is placed on terminology as a source for communication and training and the role of standardisation regarding terminological principles and practice. All the aspects related to the standardisation of language and other content resources are encapsulated in the activities of and dealt with by a technical committee of the International Organization for Standardization (ISO) and a South African mirror committee at the South African Bureau of Standards (SABS). The business of these technical committees is discussed.

Chapter 13, a contribution by Thapelo Otlogetswe, proposes three strategies of identifying multiword units from a corpus of over fifteen million words to enrich a Setswana dictionary. The study is conducted on an untagged Setswana corpus using Wordsmith Tools. The proposed strategies are: the harvesting of concordance lines, the generation of concgrams and the use of word association measures. The

three strategies have been found to be effective in the extraction of multiword units.

Chapter 14 by Hugues Steve Ndinga-Koumba-Binza and Gilles Saphou-Bivigat focuses on the dictionary basis and a few lemmatisation issues of a planned encyclopedic dictionary for Yilumbu, a developing language spoken in Gabon and in Congo. An early study on the planning of an encyclopedic dictionary for this language contains a number of discrepancies on the proposed dictionary basis and the lemma selection. In this chapter, the authors re-examine the same topics and suggest new perspectives.

In chapter 15, Rufus Gouws focuses on macrostructural and lemmatisation choices in a recently revised LSP dictionary. It is shown how different kinds of multiword and compound terms receive different types of lexicographic presentations. The main emphasis is on those sublemmata condensed to partial lemmata and presented in article niches and nests. Some of these lemma types are discussed, and innovative strategies that have been followed, e.g. the use of macrostructurally-isolated and double-layered sublemmata, are analysed. Using a contemplative approach, this paper endeavours to contribute to the expansion of a theoretical model by making the procedures that are discussed accessible to future lexicographers.

Section four also comprises five chapters. In chapter 16, Cindy McKellar and Hendrik Groenewald show that the utilisation of frequency-based data selection techniques for the generation of training data results in an increased learning rate when applied to statistical machine translation systems for three resource-scarce South African languages, namely Afrikaans, isiZulu and Sepedi. The paper first gives a general introduction to statistical machine translation. This is followed by an overview of related work and frequency-based data selection. The final part of the contribution gives information about the experimental setup and presents the results of the various machine translation systems. An interpretation of the results is provided together with some directions for future work.

Chapter 17 by Gerhard van Huyssteen, Aditi Sharma Grover and Karen Calteaux aims at getting a better understanding of business and design issues related to interactive voice responses (IVRs) in a multilingual, emerging market such as South Africa, in order to shed light on the challenges relating to voice user interface (VUI) design for such markets. It is an attempt to provide a first snapshot of the situation in South Africa, and to explicate some of the challenges. Among the 34 selected South African IVRs investigated, only 9 were found to have a multilingual offering, with only 5 having some form of speech input. Cost is the major driver for multilingual IVRs overshadowing the many positive business drivers in support of multilingual IVRs.

In chapter 18, Febe de Wet, Thomas Niesler and Christa van der Walt describe quantitative indicators obtained from undergraduate university students, with a view to automatically assess their oral proficiency. By applying automatic speech recognition methods to the automatic assessment of oral proficiency and listening comprehension, these logistical difficulties can be alleviated. Results of an automatic test are compared with human evaluations of the same data, as well as with the results of written placement tests, to determine the relationship between these approaches.

Chapter 19 is a contribution by Henning Bergenholtz who focuses on concepts for different types of information tools. The author indicates that occupation of lexicography by linguistics has serious adverse effects on the theoretical development of metalexicography and the physical products of dictionary work. In particular, the British tradition in which lexicography is classified as a field without theory and for which the linguistic theories should suffice is sharply criticised. The German tradition, while arguing at a high theoretic level, also uses the linguistic theories and forms of presenting research. The author then argues that the view that dictionaries must of necessity document extensively in the form of polyfunctional dictionaries is part and parcel of this traditional perspective. This does not apply to printed dictionaries only, but allegedly also – or even specifically – to internet dictionaries. The counterthesis postulated in this contribution is that the internet makes it even more possible than printed dictionaries to produce monofunctional dictionaries which, depending on their function, can be derived from one and the same database.

Lastly, chapter 20 is another contribution by three colleagues based in three different countries including Germany, Poland and Nigeria. In this chapter Daffyd Gibbon, Ugonna Duruibe and Jolanta Bachan present a tutorial approach to speech technology infrastructure development for a less resourced Niger-Congo tone language (Igbo), with well-known components but a new integrative strategy for creating a prototype digital signal processing front-end for a Text-to-Speech synthesiser. Text parsing problems are only dealt with in passing because the main focus is on developing a synthetic voice for a restricted practical scenario, a market information system. A generic but practical strategy for training non-specialist personnel with linguistic and/or computational skills based on a restricted domain lexicon, Finite State techniques and a traditional rule-based diphone synthesis method are described.

It is indeed a pleasure to pay tribute to the authors and co-authors who contributed to this Festschrift as well as to the referees of the papers that were submitted. Appreciation is also herein acknowledged for the firm support of Professor Wannie Carstens (Director: School of Languages, North-West University, Potchefstroom Campus), Professor Hein Viljoen (former Director: Research Unit for Language and

Literature in the South African Context, North-West University, Potchefstroom Campus), Mrs Elsa van Tonder and the entire staff of the Centre for Text Technology (CTexT). They carried this project with us from the moment they were told about it to the final stages of the publication of this book. They also managed to prevent information about this project from reaching Justus Roux's ears during various conversations. The project received funding from the Research Unit for Language and Literature in the South African Context, Potchefstroom Campus, North-West University.

Finally, honouring Professor Justus Christiaan Roux through this volume is an immense pleasure for both the editors and the contributors.

**H Steve NDINGA-KOUMBA-BINZA**  
**Sonja E. BOSCH**

---

# **PART 1:**

## PHONETICS AND PHONOLOGY

---

## CHAPTER 1

# STRESS ASSIGNMENT IN BLACK SOUTH AFRICAN ENGLISH<sup>1</sup>

Sabine Zerbian

*University of Potsdam, Potsdam, Germany*  
sabine.zerbian@uni-potsdam.de

### 1. INTRODUCTION

West-Germanic languages like English, German or Afrikaans are stress languages in which in any given polysyllabic lexical content word, such as a verb or a noun, one syllable carries the main stress. Stress is expressed acoustically by higher pitch, longer duration and/or higher intensity. In West-Germanic languages, the location of the stressed syllable within the word is not restricted to a specific syllable as it is in languages like French (last syllable) or Polish (penultimate syllable). The placement of stress in English is mainly determined by syllable weight and morpho(-syntactic) information (Giegerich 1992). There are only very few minimal pairs which are differentiated in their meaning by stress placement alone, such as 'differ and de'fer for many speakers or noun-verb pairs such as 'abstract and ab'stract (where a high colon appears before the stressed syllable).

Not all languages show the word-prosodic system of stress. In South African tone languages, for example, each syllable of a word carries a specific tone, either high or low. Tone is expressed acoustically by high pitch (not by increased duration), thus manipulating one of the same acoustic parameters as stress. However, in contrast to stress languages, the tonal specification of a syllable is determined lexically and/or morphologically. As a result, a word can have several high tones. Tone frequently differentiates between the meanings of words, e.g. Sotho *bóna* – 'see' versus *boná* – 'they', where an acute accent refers to a high tone and a low tone is left unmarked.

It has been reported in the literature that the West-Germanic fixed stress system poses a challenge in language contact. This holds for the acquisition of word stress in learner varieties (Dupoux *et al.* 1997; Dupoux *et al.* 2007) as well as for the establishment of word-prosodic systems in varieties emerging from genuine language contact, such as in bilingual speakers (Dupoux *et al.* 2009). Similarly, speakers of a language with a different word-prosodic system, such as a tone system, as well as speakers of a stress language whose stress assignment differs from the target language face difficulties with the rules that govern stress placement in a specific language (Peperkamp *et al.* 2010). The authors argue for

differences in the phonological representation of the first and target languages as the cause for the difficulties.

Against this background, it is not surprising that it has been repeatedly stated in the literature that one of the striking linguistic features of the contact variety Black South African English (BSAE; for a thorough discussion of this problematic term see Da Silva 2008:98f.) is stress placement at the word level that differs from that in General South African English (Lanham 1984, Wright 1996, De Klerk & Gough 2002).

The anecdotal evidence remained common “knowledge” until it was more thoroughly investigated in work by Van der Pas *et al.* (2000) and van Rooy (2002). Both studies reach the conclusion that word stress in Black South African English/Tswana English<sup>2</sup> is less deviant from the Standard English target patterns than commonly believed. Whereas Van der Pas *et al.*’s (2000) analysis argues for speaker-dependent coherent stress systems, van Rooy’s (2002) work makes the strong claim that there are indicators of a single stable system for stress assignment in Tswana English across speakers.

The current contribution tests the claim by van Rooy (2002) of a speaker-independent coherent system of stress assignment in Black South African English against new empirical data. It thus tests the predictions of one theory by applying it to data from a new corpus. It honours Prof. Justus C. Roux by relating to several of his research interests, both in content and methodology. It particularly deals with suprasegmental information at the word-level, which has been the topic of his work on the Nguni languages (e.g. Roux 1995, 1998). The chapter furthermore concentrates on the reflexes in Black South African English, another focus of Prof. Roux’s work (cf. Roux & Louw 2001; Roux *et al.* 2005).

The remainder of the chapter is organised as follows: Section 2 presents the background to stress in BSAE and introduces van Rooy’s (2002) algorithm for stress assignment in this variety. Then, van Rooy’s (2002) algorithm is tested by comparing its predictions against a corpus of data which is described in section 3. The results are presented and discussed in section 4. Section 5 discusses the findings with respect to their implications for the phonological representation of stress in BSAE and with respect to the limitations of control over usage frequency in corpus data. The paper concludes in section 6.

## 2. DO WE NEED LINGUISTIC THEORY? RANDOMNESS AND SYSTEMATICITY

Previous scholars entertained the view that word stress in BSAE is “*assigned idiosyncratically, very often on the penultimate syllable, following the phonological rule in Bantu languages where this syllable is lengthened*” (de Klerk & Gough 2002:361; see also Hundleby 1964:80-81). In their experience “*BSAE-speaking students of linguistics*

indicate a very marginal ability to assign native-speaker stress pattern to words" (de Klerk & Gough 2002:361.).

Van Rooy (2002) developed two hypotheses from this position: the penult-hypothesis and the random-hypothesis. According to the penult-hypothesis, stress in BSAE is attracted to the penultimate position of a word. The preference for the penultimate position has been claimed to be due to the fact that in the South African Bantu languages, the penultimate syllable of a word is predictably lengthened. According to the random-hypothesis, stress in BSAE is assigned randomly because there is no coherent stress system in this variety of English. In his work, van Rooy (2002) argues that neither of the hypotheses can account for his data in a satisfactory way. Instead, he suggests an analysis within the theoretical framework of Optimality Theory (OT) (Prince & Smolensky 1993) which is best suited to account for the observable variation in stress placement. His analysis will be rephrased here, without going into the technical details of OT.

Starting out from the basic requirement that all lexical content words need to have one primary stress, van Rooy (2002) isolates three factors that account for almost all of the data. The first is that stress falls on the penultimate syllable, as in (1a, b). This observation is central in the penult-hypothesis, and explanatory reference is made to the prosodic system of South African Bantu languages. However, the preference for the penultimate syllable is a property of varieties of English in general and results from the tendency of English stress to fall on the right edge of the word, with the caveat that the final syllable is generally exempt from stress assignment, especially in words with more than two syllables (Chomsky & Halle 1968).

<b>(1) Standard English</b>	<b>Black South African English</b>
a. 'conduct	['kan.dat]
b. 'seventies	[se.'ven.tis]
c. re'lax	[ri.'leks]
d. de'bate	[di.'bet] (underlying [eɪ] diphthong in <i>debate</i> )

However, stress in BSAE does not always fall on the penultimate syllable, as seen in (1c, d), and these exceptions are not predicted by the penult-hypothesis. Van Rooy (2002) observes that in BSAE the phonological shape of the final syllable also determines stress assignment, more specifically either a consonant cluster in the coda of the final syllable (thus syllable weight; see 1c) or the presence of an underlying diphthong in the final syllable (see 1d) leads to such syllable receiving stress. For the consonant cluster the surface form is decisive, i.e. the phonetic output that a given speaker produces, not the underlying form. Van Rooy (2007) describes how the underlying and surface form of consonant clusters might differ in BSAE. However, for the diphthong + consonant sequence it is the underlying



form that counts. This is illustrated in (1d) by the vowel [e] as a BSAE allophone of the underlying diphthong /eɪ/ (as in *pay*) which attracts stress to the final syllable. Van Rooy (2002:151) summarises both observations in characterising the BSAE stress system as quantity-sensitive so that superheavy final syllables attract stress, where superheavy refers to syllables which contain either a vowel and two consonants in the coda, or an allophone of a diphthong and a coda consonant.

These generalisations of stress assignment in BSAE are said to apply to both morphologically simple and morphologically complex words. However, for the latter one needs to differentiate between two kinds of suffixes (or affixes more generally): on the one hand those which are incorporated into the stem for the purposes of stress assignment, called opaque suffixes by van Rooy (2002:153), and on the other hand those whose presence is ignored in stress assignment, called transparent suffixes.

The basis for the categorisation of a given affix as either opaque or transparent is somewhat unclear. Van Rooy (2002:154) seems to suspect a division along derivational (=opaque) and inflectional (=transparent) affixes, and points out that “the ostensibly transparent third person [-s] and past tense/participle [-d] suffixes of English verbs” actually follow the analysis of opaque suffixes. Other examples of opaque suffixes are the nominal suffixes *-ment*, *-ion*, *-or* and *-ature* as well as the adjectival suffixes *-ant* and *-able*. Examples of transparent suffixes are the gerund affix *-ing*, plural *-s* and comparative *-er*.

Van Rooy's algorithm allows clear predictions as to which stress pattern one would expect for a given word in BSAE, provided that the underlying form (in the case of the diphthong /eɪ/), the surface segmental structure (for the coda clusters), and the morphological structure are known.

### 3. REPRODUCIBILITY: TESTING THE PREDICTIONS OF THE STRESS ALGORITHM

In order to test if the predictions of van Rooy's (2002) algorithm are borne out in new data, the stress patterns of polysyllabic content words in an already existing corpus of spoken Black South African English have been analysed.

The corpus contained the speech of 14 speakers of Black South African English. All participants were students at the University of the Witwatersrand, Johannesburg, aged between 19 and 30. Based on the phonological features of their speech and their performance in an English test (Quick Placement Test of English; QPT) they can be described as speakers of the mesolect (cf. van Rooy 2002). They all reported an African Bantu language as their first language and most gave English as or among their preferred languages (8/14). They obtained an average QPT score of 62/100.

As for the pronunciation features of the participants' speech, the speakers showed phonological features typically reported for Black South African English (van Rooy 2004), e.g. mergers in vowel quality, a trilled /r/ (predominantly male speakers; cf. Hartmann & Zerbian 2009), and overall rhythm (Coetzee & Wissing 2007).

Speakers were asked to read nine short paragraphs, containing between 35 and 65 words each. These paragraphs each constitute a brief coherent story. The speakers had time to familiarise themselves with the paragraphs before their speech was recorded. The nine paragraphs contained 79 polysyllabic content words which were subsequently transcribed with respect to their segmental and suprasegmental features by the author of the study and a student researcher. Particular attention was paid to stress realisation and surface consonant deletions<sup>3</sup>.

The corpus thus consisted of 79 target words by 14 speakers, rendering 1106 possible tokens. Due to some speaker hesitations the actual number of tokens was 1099. Table 1 in the appendix gives an exhaustive list and detailed breakdown of the target words and their approximate realisations, organised according to linguistic parameters such as number of syllables and morphological complexity.

#### **4. BEYOND IMPRESSIONS: ANALYSIS AND DISCUSSION**

##### **4.1. Supporting the Stress Algorithm**

Table 1 in the appendix presents the realisation of polysyllabic content words by 14 mesolect speakers of South African English. The first observation when comparing the actual realisations to the General South African English realisation is that the divergences between the two varieties in the segmental domain are considerable (mostly with respect to vowel quality and the manner of articulation of the r-sound). However, there is a just as well considerable agreement in the location of word stress. Thus, both Van der Pas *et al.*'s (2000) and van Rooy's (2002) observations are confirmed, namely that there is more interspeaker agreement in the placement of word stress in BSAE than anecdotally suggested. Furthermore, van Rooy's (2002) algorithm is confirmed in a variety of ways that will be presented and discussed in detail below.

The fact that most words are realised with stress on the penultimate syllable is accounted for by the important generalisation in van Rooy's (2002) algorithm, namely that stress is usually on the penultimate syllable. The algorithm's rule that superheavy syllables attract stress to the final syllable correctly predicts and accounts for the realisation of final stress in examples such as in (2). In both cases, an underlying diphthong in the final syllable paired with a stem-final consonant leads to a superheavy syllable (VVC) which attracts stress. (Note that the final

syllables in (2) are superheavy even without the Present Tense -s or the Past Tense -d.)

<b>(2)</b>	<b>orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
	arrived	əraɪvd	a'raɪvd	a'raɪv, a'raɪvd (7), a'laɪvd (3), ə'raɪvd, ə'laɪvd, a'raɪvd
	complains	kəmpleɪns	kom'plens	kəm'plens (12), kʌm'plens, 'kəmplens

The examples in (3) confirm the claimed status of the Past Tense suffix -d as an opaque suffix which is incorporated into the stem for the purposes of stress assignment. This is because it is only by the addition of the suffix that the last syllable in the examples in (3) becomes superheavy and therefore attracts stress as correctly predicted by van Rooy's algorithm. Note that only underlying diphthongs but not long vowels as in (3) constitute a long syllable nucleus in van Rooy's BSAE stress algorithm.

<b>(3)</b>	<b>orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
	divorced	dɪvɔːsd	dɪ'vosd	dɪ'vosd (2), dɪ'vɔsd (6), dɪ'vɔrs (2), dɪ'vɔrsd (3), dɪ'vɔs
	received	rɪsɪːvd	rɪ'sɪvd	rɪ'sɪvd (4), ɹɪ'seɪvd, rɪ'sɪt, ɹɪ'sɪvd (2), rɪ'zɪv, rɪ'zɪvd (5)

The example in (4) supports van Rooy's assumption that not all suffixes behave similarly with respect to stress assignment. The plurals -s in (4) would render the final syllable superheavy (i.e. VCC). It would thus attract stress were it not for the status of plural -s as a transparent suffix which is ignored for the purposes of stress assignment.

<b>(4)</b>	<b>orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
	lions	laɪəns	'laɪens	'laɪons (7), 'laɪəns (7)

Van Rooy (2002) claims that in order for a syllable to be superheavy and thus to attract stress it needs to have either a diphthong in the underlying form followed by at least a single consonant or a vowel followed by two consonants at the surface. Example (5) supports the view that it is an underlying diphthong that is at issue. On the surface, the example in (5) is pronounced with a monophthong by many

speakers, thereby creating a phonological context which does not warrant the attraction of stress.

<b>(5) orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
remote	rɪməʊt	ri'mot	ri'mot (7), ɪi'mout (6), ri'mout

The examples in (6) support van Rooy's (2002) claim that for coda consonant clusters it is the surface form of these underlying clusters that is considered when determining the syllable weight, and not the underlying form. In both cases, stress would be predicted on the final syllable when based on the underlying form given that they are superheavy (VCC). However, the majority of speakers (11 and 13 respectively) do not realise the final consonant clusters in which case the prediction is that stress falls indeed on the penultimate syllable as realised by the speakers.

<b>(6) orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
promised	prəmɪsd	pro'mist	'pɹɔmɪz (11), 'pɹɔmɪzd (2)
boyfriend	bɔɪfrend	bɔɪ'frent	'bɔɪfrɛnd, 'bɔɪfrɛn (10), 'bɔɪfrɛn (3)

#### 4.2. CHALLENGING THE STRESS ALGORITHM

As the examples in (2) – (6) show, the new data that have been collected support the BSAE stress algorithm by van Rooy (2002) in a variety of ways. However, there are also cases that are not captured by the algorithm. These will be presented and discussed now.

There are few instances in which the surface form does not predict stress assignment. The reader might have noticed these instances in (3) and (6). They are repeated in (7) for convenience. In (7a), we see some individual realisations without a final consonant cluster, so that stress should actually have been on the penultimate syllable. In (7b), the realised consonant cluster should have attracted the stress to the final syllable.

<b>(7) orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
a. divorced	dɪvɔ:sd	dɪ'vosd	dɪ'vɔs
received	rɪsi:vd	ri'sivd	ri'sit, ri'zɪv
b. promised	prəmɪsd	pro'mist	'pɹɔmɪzd (2)

As the number of realisations show, these instances are relatively few and we therefore present them for completeness' sake but do not consider them true challenges to the postulated algorithm. Other examples might prove to be more difficult to overrule. For (8), e.g., the algorithm predicts stress on the penultimate

syllable (because long vowels do not count as heavy in van Rooy's analysis), but all speakers realise the standard pattern of final stress.

<b>(8) orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
believe	bɪli:v	'biliv	bə'li:v (5), bi'li:v (9)

If underlying long vowels were to count towards a heavy syllable in the BSAE stress algorithm, final stress in this example could be predicted. Van Rooy (2002:152) notes that the restriction that only underlying diphthongs but not long vowels qualify as long syllable nuclei in his algorithm "raises some interesting questions about the nature of vowel quality in Tswana English specifically, and probably in BSAE generally. [...] No known research addresses this possibility adequately." Example (8) reiterates the need for a thorough investigation into vowel quality and vowel length in BSAE.

There is one further example which mirrors (8) and which is left unaccounted for. It is given in (9). Here stress is final although the syllable only consists of a diphthong. Again, all speakers realise the target stress pattern.

<b>(9) orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
apply	əplaɪ	'aplai	a'plai (10), ə'plai (3)

Also the opposite case occurs: The BSAE algorithm predicts final stress but stress is actually (in agreement with the standard pronunciation) realised on the penultimate syllable. Examples are given in (10).

<b>(10) orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
island	aɪlənd	aɪ'lænd	aɪ'lend, 'aɪlənd (13), 'aɪlənd
mountains	maʊntəns	maʊn'tens	'mauntens (9), maʊn'teɪns (2)
parents	peərənt	pɛ 'rɛnts	'parents, 'peɪənt, 'pɛrɛnst, 'pɛrɛnst
apartment	əpɑ:tmənt	apɑt'ment	a'patment (9), a'pɑ:tment (5)

In (10) the final syllables are superheavy due to realised consonant clusters in coda position. In all cases, speakers predominantly realise the target stress pattern.

The examples in (9) and (10) could be argued to constitute speaker-dependent idiosyncrasies or lexical exceptions. However, the examples in (11) show a principled restriction of van Rooy's stress algorithm<sup>4</sup>. For all examples in (11), stress is predicted to occur either on the final or on the penultimate syllable whereas it is actually realised on the antepenultimate syllable in agreement with the Standard English stress pattern.

(11)	<b>orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
	elephant	ɛlɛfənt	ɛlɛ'fɛnt	'ɛlɛfɛnt (14)
	hesitant	hezɪtənt	hezi'tɛnt	'hɛzətɛnt (7), 'hɛzɪtɛnt (3), hɛ'zɪstənt (2), hɛzi'tənt (2)
	January	dʒænjuəri	dʒɛn'wari	'dʒɛnuəri (4), 'dʒɛnuarə, 'dʒɛnjuəri, 'dʒanuəri (3), 'dʒɛnuar, 'dʒɛnuə, 'dʒanuəri, 'dʒanuəri, 'dʒanuəri
	holiday	hɒlədeɪ	hɒ'lɪdeɪ	'hɒlɪdeɪ (12), hɒli'deɪ (2)
	ringmaster	rɪŋmaːstə	rɪŋ'masta	'rɪŋmaːstə (11), 'rɪmaːstə, rɪŋ'maːstəs, 'rɪŋk'mastə
	everywhere	evriweə	ev'riweɪ	'evriweɪ (10), 'evriweɪ (4)

Admittedly, some of the examples show morphologically complex words which should be considered compounds and for which it has not been made explicit whether and how the algorithm works. What becomes clear, however, is that van Rooy's (2002) algorithm for stress in BSAE generally excludes the possibility of antepenultimate stress unless the last syllable is a transparent suffix, as in *travelling*. Because transparent suffixes are ignored in stress assignment, the algorithm can accurately derive antepenultimate stress in these cases. Nevertheless, speakers do realise antepenultimate stress also on morphologically simplex words, as most of the realisations in (11) show.

In addition to lending support to the algorithm and to eliciting some challenging and interesting counter-examples, the newly evaluated data also allow making a contribution to the classification of suffixes into opaque and transparent suffixes as far as stress assignment is concerned. The examples in (12) suggest that the adverb suffix *-ly* acts as a transparent suffix (which is consequently ignored for the purposes of stress assignment) so that the initial stresses in the trisyllabic words can be accounted for.

(12)	<b>orthography</b>	<b>underlying</b>	<b>predicted</b>	<b>realised</b>
	recently	riːsntli		'ɹɪsɛntli (5), 'rɪsɛntli (7), 'rɪsɛnt
	really	riːəli, rɪəli		'rɪli (7), 'ɹɪli (6), 'ɹɪəli
	willingly	wɪlɪŋli		'wɪlɪŋli (13), 'wɪlɪŋli

## 5. DISCUSSION

The previous section has presented new data from a corpus of read speech which was evaluated against the BSAE stress algorithm proposed by van Rooy (2002). It could be shown that the algorithm accounted for many of the occurring stress realisations and that the corpus thus provided support for all of the assumptions made by van Rooy (2002). It also became clear that the algorithm – not surprisingly – cannot account for all the data. The challenge seems to be the assignment of antepenultimate stress. The logical step forward would now be to propose a modification of the algorithm that would do exactly that. Instead the discussion section wants to raise two concerns that prevent me from doing so: these concerns relate to the representation of word prosody in language contact, and the role of grammar and frequency in language.

### 5.1. Word-Prosodic Systems in Contact Languages

A first concern in contributing to the postulation of a stress system in mesolect speakers of BSAE relates to the nature of word prosodic systems in contact languages. By definition, mesolect speakers use a variety which lies between the two languages involved in the language contact situation (Bickerton 1971). In this case, BSAE is between English and the South African Bantu languages. With respect to the word prosodic system the initial assumption would be that the word prosodic system of mesolect speakers lies between the stress system of English and the tone system of Bantu languages. However, developing a stress algorithm for mesolect speakers of BSAE in fact assumes a stress system, an assumption which contradicts the above-mentioned assumption. Assuming a stress system for such mesolect speakers therefore needs careful motivation.

Zonneveld (2010) interprets the near-native results of some few Tswana English participants in the study of Van der Pas *et al.* (2002) as evidence of a successful mastery of the English stress system. Three to ten out of the 50 participants in Van der Pas *et al.*'s study produced stress patterns comparable to a Canadian English control group, tested in Pater (1997). Furthermore, stress patterns produced by speakers of Tswana English were generally closer to the ones of controls than the French L1 group also tested in Pater (1997).

Nevertheless, only little is actually known about the phonological representation of prosody in contact languages. Work by Dupoux and colleagues on the perception of word stress in foreign language acquisition and bilingualism has shown that the phonological representation of a specific language's word prosodic system is very stable in language contact. Dupoux *et al.* (1997, 2007) showed that French late learners of Spanish are impaired in discrimination tasks with stimuli that vary only in the position of stress. Although both French and Spanish are stress languages,

they differ in the predictability of the position of word stress: whereas French has predictable word-final stress, Spanish shows flexible stress similar to English. Dupoux *et al.* (1997, 2007) carefully argue that French speakers lack a phonological (in their words *metalinguistic*) representation of contrastive stress which impairs them in discrimination tasks.

Even in simultaneous French-Spanish bilinguals the phonological representation of stress differs from that of monolingual speakers of the languages. Dupoux *et al.* (2009) found that the performance of simultaneous bilinguals in two memory tasks and one lexical decision task was intermediate between that of monolingual native speakers of Spanish on the one hand and French late learners of Spanish on the other hand. Thus, the phonological organisation of a language that a speaker is exposed to early in his/her life leaves a trace in adult grammars too.

Gussenhoven and Udofot (2010; see also Gut 2005) investigate sentence prosody in Nigerian English, a variety arising out of the contact between a stress language and a West African tone language. Based on the results of a perception experiment, they argue that Nigerian English sentence prosody is best modelled with tonal specifications for each syllable, including syllables that are unstressed in British English. Also here, the phonological representation of prosody – though sentence prosody and not word prosody in this case – shows traces of both languages involved in the contact situation.

Also the literature on the acoustic realisation of sentence stress suggests that contact languages can creatively develop prosodic systems which differ from the prosodic systems of both the substrate and the superstrate languages. Work on other New Englishes has claimed that acoustic parameters are used differently than in British or American English, e.g. an increased use of intensity in sentence intonation in Cameroon English (Talla Sando Ouafeu 2007) or a focus-independent use of intensity and fundamental frequency (F0) in Black South African English (Zerbian 2011). A manipulation of different acoustic parameters might, at least partly, be due to the fact that these African Englishes are varieties of English that arose in contact with African tone languages, and tone languages use F0 for lexical and grammatical distinctions. A thorough investigation of the acoustic correlates of stress in BSAE, though necessary, lies outside the scope of the current work.

To sum up, there is evidence from both psycholinguistic and phonetic studies that one cannot assume a clear-cut stress system in contact varieties such as the mesolect of BSAE, and that careful psycholinguistic and phonetic analysis is needed to investigate the issue further. Corpus studies such as the current one and its precursor (van Rooy 2002) are useful to delineate relevant parameters such as morphological complexity, underlying vowel quality and surface consonant clusters. Nevertheless, it seems necessary to collect carefully controlled data in



addition to corpus data to address the questions that have been raised in this section.

## 5.2. The Role of Grammar and Frequency in Language

A second concern that stands in the way of developing a solely grammar-based stress algorithm for BSAE relates to the finding that language learning seems to involve a combination of learning mechanisms (e.g. Carpenter 2010 for stress): some are innate mechanisms specific to language, like formal abstract grammatical principles such as those expressed in an OT analysis, and others are general cognitive mechanisms, like frequency statistics. Although both mechanisms are based on experience with language to some degree (after all, formal abstract grammatical principles evolve through exposure to language), Coetzee (2008) carefully differentiates between the two. Transferred to the case at hand, the question arises if stress in BSAE is solely assigned on the basis of an active algorithm or also due to lexical statistics. More precisely, does the algorithm need to account for the cases of antepenultimate stress in a principled way or are these cases produced because of the speakers' familiarity with the words in question due to lexical frequency? I agree with Coetzee (2008) that the phonological grammar is not "a simple projection of lexical statistics" (as stated e.g. by Hay *et al.* 2003:59). However, in order to disentangle the two general mechanisms, the well-documented influence of usage statistics needs to be controlled for in linguistic studies that address this question so that results can be interpreted as giving evidence about grammar *per se*.

Corpus studies do not allow this type of control. A frequent experimental paradigm that controls for usage statistics is the use of nonsense words, i.e. made-up words that do not exist in the language under investigation. Nonsense words have the advantage that language users do not have any prior experience with the stimuli used. This was the approach taken by Van der Pas *et al.* (2000) in their study on word stress in Tswana English. Interestingly, their results show that speakers of Tswana English often do stress the antepenultimate syllable: the majority of participants produced antepenultimate stress on 5 out of 16 words, such as *tadimet*, *kadowtet*, and *kapistratson*. This suggests that the occurrence of antepenultimate stress cannot solely be due to usage frequency but is a phenomenon that an algorithm would need to account for. Recently, Zonneveld (2010) reanalysed Van der Pas *et al.*'s data as showing evidence that speakers of Tswana English show an interim grammar in which final VC-syllables are extrametrical and thereby open up the possibility of antepenultimate stress. Furthermore, a quantity sensitive effect can be observed for prefinal syllables, thus accounting for the difference between *na-co-stracan* and *ka-ta-pes-tos*.

Unfortunately, Van der Pas *et al.* (2000) do not systematically include those linguistic variables that have been suggested by van Rooy (2002) as relevant for stress assignment in BSAE, such as vowel quality and syllable structure. Studies using nonsense words but controlling for these variables are indicated.

## 6. CONCLUSION

The current contribution addressed stress in the mesolect of Black South African English. It presented new data to test the algorithm suggested by van Rooy (2002). The findings confirmed that the algorithm can account for many of the observed stress patterns and thereby refutes, in line with van Rooy (2002), the previously held impression that stress assignment in this variety is either idiosyncratic or restricted to the penultimate syllable. The new data also allow for the classification of the adverb suffix *-ly* as a transparent suffix which is ignored for the purposes of stress assignment. It emerged from the data presented that antepenultimate stress in morphological simple words cannot be accounted for by the algorithm although it regularly occurs in the speech of mesolect speakers of BSAE.

This work refrained from further developing the algorithm to include antepenultimate stress, a prerequisite for such developing is greater clarity on the nature of the phonological representation of stress in contact languages, especially involving languages with two different word-prosodic systems such as the stress system of English and the tone system of the South African Bantu languages. Also, following these first explorative investigations into the emerging word-prosodic system of BSAE, data are now needed that controls for lexical frequency and usage statistics in order to allow making inferences about grammar *per se*.

## ENDNOTES

1. The research that went into this chapter was funded by the German Research Foundation (DFG), grant to the SFB 632 'Information structure' at the University of Potsdam. I want to thank all participants of the experiment, the School of Languages and Literature Study at the University of the Witwatersrand, Johannesburg, for hosting me during my research stay in September 2010, as well as the research assistants Steven Fielding, Svenja Schürmann, and Eric Tabbert for their help in gathering the data and preparing them for analysis.
2. Both studies investigate stress in black speakers with Tswana as an L1. In this paper the general cover term Black South African English is used, following several other publications (e.g. van Rooy 2004) which do not make a difference between the varieties according to the L1 (but see also Da Silva 2008:98f).
3. Vowel quality was carefully transcribed in van Rooy's (2002) study. However, because surface vowel quality or quantity has not been shown to impact on stress

assignment, only approximate vowel qualities are given in our transcriptions. The following conventions have been followed: [a] represents the low vowel independent of actual quality; it is transcribed as [ə] in underlying unstressed positions. We differentiated between [e] and [ɛ] only auditorily. Vowel lengthening is only transcribed if it was very striking.

4. A reviewer cautions against extending the predictions of van Rooy's algorithm to words with more than two syllables as his work was felt not to contain many polysyllabic words with more than two syllables. However, a count shows that 14 cases of polysyllabic words with more than two syllables are reported in van Rooy (2002).

## REFERENCES

- Bickerton, D. 1971. Inherent variability and variable rules. *Foundations of Language*, 7:457-92.
- Carpenter, A. 2010. A naturalness bias in learning stress. *Phonology* 27(3):345-392.
- Chomsky, N. & M. Halle. 1968. *The Sound Patterns of English*. New York: Harper & Row.
- Coetzee, A. W. 2008. Grammaticality and ungrammaticality in phonology. *Language* 84(2):218-257.
- Coetzee, A.W. & D. Wissing. 2007. Global and local durational properties in three varieties of South African English. *The Linguistic Review* 24(2-3):263-290.
- Da Silva, A.B. 2008. *South African English: A sociolinguistic investigation of an emerging variety*. Unpublished PhD thesis, University of the Witwatersrand, Johannesburg.
- De Klerk, V. & D. Gough. 2002. Black South African English. *Language in South Africa*, edited by R. Mesthrie. Cambridge: Cambridge University Press. 356-378.
- Dupoux, E., C. Pallier, N. Sebastián & J. Mehler. 1997. A destressing 'deafness' in French? *Journal of Memory and Language* 36:406-421.
- Dupoux, E., S. Peperkamp & N. Sebastián-Gallés. 2009. Limits on bilingualism revisited: Stress 'deafness' in simultaneous French-Spanish bilinguals. *Cognition* 114(2):266-275.
- Dupoux, E., N. Sebastián-Gallés, E. Navarrete & S. Peperkamp. 2007. Persistent stress 'deafness': The case of French learners of Spanish. *Cognition* 106(2):682-706.

- Giegerich, H.J. 1992. *English phonology: An introduction*. Cambridge: Cambridge University Press.
- Gussenhoven, C. & I. Udofot. 2010. Word melodies vs. pitch accents: A perceptual evaluation of terracing contours in British and Nigerian English. *Proceedings of Speech Prosody 2010*, 100015:1-4.
- Gut, U. 2005. Nigerian English prosody. *English World-Wide* 26(2):153-177.
- Hartmann, D. & S. Zerbian. 2009. Rhoticity in Black South African English - A sociolinguistic study. *Southern African Linguistics and Applied Language Studies* 27(2):135-148.
- Hay, J., J.B. Pierrehumbert & M.E. Beckman. 2003. Speech perception, well-formedness and the statistics of the lexicon. *Phonetic Interpretation: Papers in Laboratory Phonology 6*, edited by J. Local, R. Ogden & R. Temple. Cambridge: Cambridge University Press. 58-74.
- Lanham, L.W. 1984. Stress and intonation and the intelligibility of South African Black English. *African Studies* 43:217-30.
- Pater, J. 1997. Metrical parameter missetting in second language acquisition. *Focus on phonological acquisition*, edited by S.J. Hannahs & M Young-Scholten. Amsterdam: Benjamins. 235-261.
- Peperkamp, S., I. Vendelin & E. Dupoux. 2010. Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics* 38:422-430.
- Prince, A. & P. Smolensky. 1993. Optimality Theory: Constraint Interaction in Generative Grammar. Rutgers University Center for Cognitive Science Technical Report 2.
- Quick Placement Test of English*. 2004. Oxford: Oxford University Press.
- Roux, J.C. 1995. Prosodic data and phonological analyses in Zulu and Xhosa. *South African Journal of African Languages* 15(1):1-10.
- Roux, J.C. 1998. Xhosa: A tone or pitch-accent language? *South African Journal of Linguistics*, Supplement 36:33-50.
- Roux, J.C., J.A. Du Preez & E. De Villiers. 2005. Accent variation in South African English: Challenges for speech recognition systems. *Proceedings of the Second Language and Technology Conference*, Poznan, Poland. 15-19.
- Roux, J.C. & P.H. Louw. 2000. Black South African English (BSAE) and Speech Technology applications. *South African Journal of Linguistics*, Supplement 38:4-13.

- Talla Sando Ouaféu, Y. 2007. Intonational marking of new and given information in Cameroon English. *English World-Wide* 28(2):187-199.
- Van der Pas, B., D. Wissing & W. Zonneveld. 2000. Parameter resetting in metrical phonology: the case of Setswana and English. *South African Journal of Linguistics*, Supplement 38:55-88.
- Van Rooy, B. 2002. Stress placement in Tswana English: the makings of a coherent system. *World Englishes* 21(1):146-160.
- Van Rooy, B. 2004. Black South African English – phonology. *Handbook of Varieties of English*, vol. 1, *Phonology*, edited by E. Schneider, K. Burridge, B. Kortmann, R. Mesthrie & C. Upton. Berlin: Mouton de Gruyter. 943-952.
- Van Rooy, B. 2007. Consonant clusters and resyllabification in Black South African English. *Language Matters* 38(1):26-45.
- Wright, L. 1996. The standardization question and Black South African English. *Focus on South Africa*, edited by V. de Klerk. Amsterdam: John Benjamins. 149-162.
- Zerbian, S. 2011. Intensity in narrow focus across varieties of South African English. *Proceedings of the 17th International Congress of Phonetic Sciences 17-21 August 2011 Hong Kong*, edited by Wai-Sum Lee & Eric Zee. Hong Kong: City University of Hong Kong. 2268-2271.
- Zonneveld, W. 2010. Default, non-default, markedness and complexity in the L2 English word stress competence of L1 speakers of Setswana. *Southern African Linguistics and Applied Language Studies* 28(4):375-391

## APPENDIX

Table 1: Target words, their underlying form, their predicted output form based on van Rooy (2002, 2004) and their approximate actual realisations

Target words	Underlying form	Prediction (van Rooy 2002)	Realisations (where N≠1 is given in brackets)
<b><i>Disyllabic, morphologically simplex</i></b>			
Cindy	sɪndi	'sindi	'sindi (2), 'sɪndi (10)
Nellie	nɛli	'nɛli	'nɛli (14)
circus	sɜ:kəs	'sekas	'sœkəs (5), 'sekəs (5), 'sekas(3), 'sekəs
Lilly	lɪli	'lɪli	'lɪli (10), 'lɪli (4)
William		'wɪljɛ m	'wɪljəm (2), 'wɪljəm (2), 'wɪljəm (6), 'wɪljəm (4)
hurry	hʌɹi	'hari	'hʌɹi (6), 'hʌɹi (7)
Peter		'pɪtə	'pɪtə (12), 'petə, 'pɪdə
Sarah		'sara	'sara (5), 'sɛ ɹa (7), 'saɹa (2)
Durban		'dɛben	'dœbən (9), 'dɛbən (4), 'dɛɹbən
believe	bɪli:v	'bɪliv	bə'li:v (5), bɪ'li:v (9)
busy	bɪzi	'bɪzi	'bɪzi (10), 'bɪzi (4)
China	tʃaɪnə	ʔʃaɪna	ʔʃaɪna (11), 'ʃaɪna (2), ʔʃaɪn
English	ɪŋɡlɪʃ	'ɪŋlɪʃ	'ɪŋlɪʃ, 'ɪŋlɪʃ (13)
remote	rɪməʊt	'rɪmot	ri'mot (7), ɹi'mout (6), ri'mout
island	aɪlənd	aɪ'lend	'aɪlənd (13), 'aɪslend
very	veri	'veri	'vɛɹi (6), 'vɛɹi (8)
letter	letə(r)	'letə	'letə (11), 'let (2)
partner	pɑ:tənə(r)	'patna	'patnə (9), 'pa:tne (4), 'pa:tne

jealous	dʒeləs	'dʒelas	'dʒelas (6), 'dʒeləs (7), 'kjelas
Amy		'emi	'eimi (12), 'emi (2)
visit	vɪzɪt	'vizit	'vɪzɪt (13), 'vɪsɪt
mother	mʌðə(r)	'maðə	'madə (4), 'maðə (4), 'maðe (6)
Mary		'meri	'mæri (3), 'mɛɪ (8), 'mɛri (2), 'mæɹə
Thabo		'tabo	'tabo (14)
every	evri	'evri	'ɛvri (7), 'ɛvɪɪ (6)
soccer	sɒkə(r)	'soka	'sɒkə (13), 'sokə
lady	leɪdi	'ledi	'ledi, 'leɪdi (10), 'led
Emma		'ema	'ɛma (14)
heavy	hevi	'hevi	'hɛvi (14)
apply	əplai	'aplaɪ	a'plai (10), ə'plai (3)
able	eɪbl	'eɪbl	'eɪbəl (14)
<b><i>Disyllabic, morphologically complex due to opaque affixes</i></b>			
wonders	wʌndə(r)s	'wondəs	'wɒndəs (7), 'wɒndəs (7)
married	mæɪɪd	'merid	'mɛrid (5), 'mɛɪɪd (8), 'mɛɪɪ
divorced	dɪvɔ:sd	di'vosd	di'vosd (2), di'vɔsd (6), di'vɔrs (2), di'vɔrsd (3), di'vɔs
received	rɪsi:vd	ri'sivd	ri'sivd (4), ɪi'sevd, ri'sit, ɪi'sivd (2), ri'zɪv, ri'zɪvd (5)
question	kwestʃən	'kwɛstjən	'kwɛʃn (2), 'kwɛʃən (9), 'kwɔɛʃn, 'kwɛstjɪn (2)
promised	prɒmɪsd	pro'mɪst	'pɹɒmɪz (11), 'pɹɒmɪzd (2)
arrived	əraɪvd	a'raɪvd	a'raɪv, a'raɪvd (7), a'ɹaɪvd (3), ə'raɪvd, ə'ɹaɪvd, a'raɪvd
unwell	ʌnwel	'anwɛl	'anwɛl (4), an'wɛl (10)
complains	kəmpleɪns	kom'plɛns	kɒm'plɛns (12), kʌm'plɛns, 'kɒmplɛns

loses	lu:zɪz	'lu:zɪz	'lu:zɪz (14)
<b>Disyllabic, morphologically complex due to transparent suffixes</b>			
slipper	slɪpə	'slɪpə	'slɪpə (11), 'slɪpə (3)
owner	əʊnə	'ona	'onə (14)
mountains	maʊntəns	maʊn'tens	'maʊntəns (9), maʊn'teɪns (2), 'maʊnten (2), 'maʊntes
looking	lʊkɪŋ	'lʊkɪŋ	'lʊkɪŋ (14)
feeling	fi:lɪŋ	'fɪlɪŋ	'fɪlɪŋ (14)
planning	plænɪŋ	'plænɪŋ	'plænɪŋ (13), 'pɛnɪŋ
lions	laɪəns		'laɪons (7), 'laɪəns (7)
tamers	teɪməs	'teɪmas	'teɪməs (10), 'teɪməs (2), 'təməs, 'tɛməs
parents	pɛərənts	pɛ'rents	'parents, 'parens (5), 'pə:ɛns (2), 'pɛ:ɛnt, 'pɛ:ɛnst, 'pɛ:ɛns, 'pɛ:ɛnst, pɛ':ɛns (2)
watching	wɒtʃɪŋ	'wɒtʃɪŋ	'wɒtʃɪŋ (12), 'wɒtʃ (2)
better	bɛtə	'bɛtə	'bɛdɛ (3), 'bɛtɛ (5), 'bɛtɛ (2), 'bætɛ (3), 'bæt
going	ɡəʊɪŋ	'ɡoɪŋ	'ɡoɪŋ (2), 'ɡoʊɪŋ (11), 'ɡoɪn
boxes	bɒksɪs	'bɒksɪs	'bɒksɪs (14)
moving	mu:vɪŋ	'mu:vɪŋ	'mu:vɪŋ (13)
<b>Disyllabic compounds (with transparent suffixes)</b>			
birthday	bɜ:θdeɪ	'beθdeɪ	'bɜ:θdeɪ (13), bɜ:θ'deɪ
farewells	feəwɛls	'fɛwɛls	fɛ'wɛls (4), 'fɛwɛls (9), fɛ'wɛl
upset	ʌpset	'apset	'apset (4), ap'set (10)
boyfriend	bɔɪfrend	bɔɪ'frent	'bɔɪfrend, 'bɔɪfren (10), 'bɔɪfɛn (3)
passport	pɑ:spɔ:t	'paspot	'paspɔ:t (10), 'pɑ:spɔ:t (3), 'paspɔt
<b>Trisyllabic</b>			
another	ənʌðə	a'naðə	a'nada (2), a'naðɛ (8), a'naðə, ə'naðə (3)



elephant	ɛləfənt	ɛlə'fɛnt	'ɛləfɛnt (14)
banana	bənana	bə'nana	bə'nana (12), bə'nana (2)
hesitant	hezɪtənt	hezi'tɛnt	'hezətɛnt (7), 'hezɪtɛnt (3), he'zɪstənt (2), hezi'tənt (2)
January	dʒænjuəri	dʒen'wari	'dʒɛnuəri (4), 'dʒɛnuərə, 'dʒɛnjuəri, 'dʒanuəri (3), 'dʒɛnuər, 'dʒɛnuə , 'dʒanuəri, 'dʒanuəri, 'dʒanuəri
<b>Trisyllabic, morphologically complex with opaque suffixes</b>			
commences	kəmensɪs	kə'mɛnsɪs	kə'mɛnsɪs (12), kə'mɛns (2)
recession	rɪseʃn	rɪ'seʃɛn	rɪ'seʃən (11), rɪ'seʃɛn (2), rə'seʃən
apartment	əpa:tmənt	apət'mɛnt	a'pətmɛnt (9), a'pə:tmɛnt (5)
<b>Trisyllabic, morphologically complex with transparent suffixes</b>			
inviting	ɪnvartɪŋ	ɪn'vartɪŋ	ɪn'vartɪŋ (14)
travelling	trævəlɪŋ	'trævelɪŋ	'trævelɪŋ (12), 'trævelɪn (1), 'træveli (1)
carrying	kæərɪŋ	'kɛərɪŋ	'kɛrɪ'ŋ (6), 'kɛɪɪŋ (2), 'kɛrɪŋ (6)
<b>Trisyllabic, morphologically complex compounds</b>			
ringmaster	rɪŋma:stə	rɪŋ'masta	'rɪŋma:stə/ə (11), 'rɪma:stə, rɪŋ'ma:stəs, 'rɪŋk'mastə
holiday	hɒlədeɪ	hə'lide	'hɒlɪdeɪ (12), hɒlɪ'deɪ (2)
everywhere	evriweə	ev'riwe	'evɪɪwe (10), 'evɪɪwe (4)
<b>Suffixes unclassified</b>			
mysterious	mɪstɪəriəs	mɪ'sti:əs	mɪs'təriəs (2), mɪ'sti:əs (2), mɪ'sti:əs (2), mɪ'sti:əs (2), 'mɛstriəs, mɪ'stəriəs, mɪs'tɪrɪəs, mɪs'tərəəs, 'mɪstɪəəs, 'mɪstrəəs,
recently	rɪ:sntli		'ɪsɛntli (5), 'rɪsɛntli (7), 'rɪsɛnt
really	rɪ:əli, rɪəli		'rɪli (7), 'ɪli (6), 'ɪəli
willingly	wɪlɪŋli		'wɪ lɪ ŋli (13), 'wɪlɪ ŋli

## CHAPTER 2

# **A COMPARATIVE STUDY OF EXPLOSIVE [b] AND IMPLOSIVE [ɓ] IN NGUNI**

Shamila Naidoo

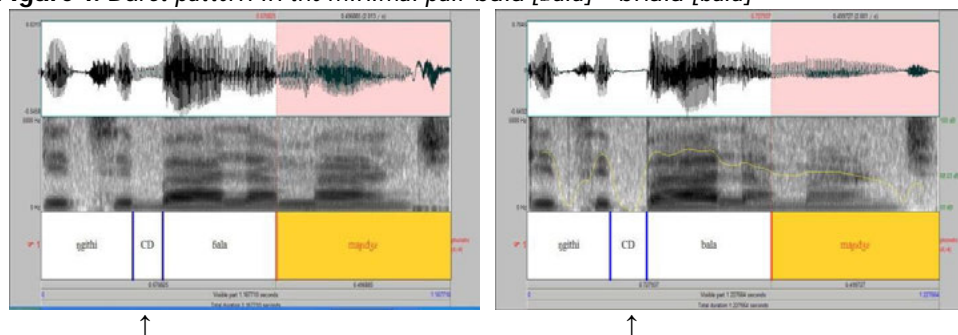
*School of IsiZulu, University of Kwa-Zulu Natal, Durban, South Africa*  
naidoosh@ukzn.ac.za

### **1. INTRODUCTION**

According to Clements and Rialland (2005:21) the implosive<sup>1</sup> is 'a characteristic feature of broad areas of Africa'. It is found among languages from the Sudanic belt where the occurrence of **ɓ** and **ɗ** is twelve times more frequent than elsewhere. The Cushitic and Omotic languages of the east zone and the Bantu languages of the south zone are also home to the implosive. The combination of rare occurrence and complex phonetic quality of the implosive has led to experimental and impressionistic investigations of this sound<sup>2</sup>. Naidoo (2010) conducted an experimental study of the Zulu implosive. That study identified three points<sup>3</sup> which are intrinsic to classifying the implosive [ɓ] and explosive [b] as discrete sounds:

- i. The presence vs absence of a prominent burst on the spectrograms of explosive vs implosive sounds
- ii. Differences in closure duration (CD) values with the explosive having a longer CD value than the implosive
- iii. Lower burst amplitude measurement for the implosive

Figure 1 is used to illustrate the qualitative aspect of this experimental investigation. Intrinsic to differentiating between an implosive and explosive sound, is the burst pattern. Figure 1 contrasts the spectrograms of words containing the implosive [ɓ] and explosive [b]. The absence of a burst on the spectrogram of *bala*, which contains the implosive [ɓ], is consistent with the position of Ladefoged and Maddieson (1996:82) who state that with implosives 'the stop burst is less evident'. Johnson (1997:133) corroborates this, explaining that 'the stop release bursts of implosives (glottalic ingressive sounds), are weaker than those of pulmonic stops'.

**Figure 1:** Burst pattern in the minimal pair *bala* [bala] – *bhala* [bala]

Also, qualitatively evident on Figure 1 is a difference in closure duration (CD) values. These qualitative assessments, in conjunction with quantification of the CD values and Intensity measurements, form the basis of the Naidoo (2010) experimental study. The latter was a catalyst for this chapter in which the first stage results<sup>4</sup> of a comparative experimental study of the Nguni implosive and explosive sounds are presented. The focus of the study is to determine whether or not the explosive [b] and implosive [ɓ] present the same results, across the three languages, in terms of the parameters burst amplitude and closure duration (CD). According to my knowledge there has been no such comparative experimental study.

This chapter is broadly divided into four parts.

- The classification of the implosive in the Nguni languages is briefly discussed.
- The procedure for the experimental investigation is explained.
- The results are presented and discussed.
- Some conclusions are presented.

## 2. THE NGUNI IMPLOSIVE [ɓ]

The status of the Nguni implosive [ɓ] is not a controversial issue. Certainly, the Zulu implosive [ɓ] has been the subject of investigation<sup>5</sup> but not so for the Swati and Xhosa implosive [ɓ] sounds. Recent phonetic textbooks all acknowledge the existence of the implosive [ɓ] in the Nguni languages. Taljaard and Snyman (1991a:68), Finlayson *et al.* (1993:73) and Taljaard and Snyman (1991b:68) include the implosive [ɓ] in the inventories of Swati, Xhosa and Zulu, respectively.

## 2.1. The Experiment: an Analysis of the Nguni Implosive [ɓ] and Explosive [b]

### Data

It was not possible to use Doke's (1926) corpus of minimal pairs<sup>6</sup> as these words did not obtain across the Nguni languages. Furthermore, it was rather difficult to find identical minimal pairs in Swati, Xhosa and Zulu. On account of this challenge, it was decided to select individual words, as opposed to minimal pairs exclusively, which occur in all the Nguni languages. These are shown in Table 1.

**Table 1: Wordlist 1**

Nguni words		Phonetic
bala	<i>count</i>	[ɓala]
bhala	<i>write</i>	[bala]
bhonga	<i>roar</i>	[bɔŋa]
bonga	<i>praise</i>	[ɓɔŋga]
buza	<i>ask</i>	[ɓuza]
bhubha	<i>perish</i>	[buba]

The preparation of the data for the analysis was done in the following phases:

- Two word lists were prepared. Word list 1, shown in Table 1 above, contained minimal pairs and individual words that occur in all three Nguni languages. In wordlist 2, the order of the words was randomised and several distracters were included. This is shown in Table 2.

**Table 2: Wordlist 2**

		Phonetic		Phonetic	
sala	<i>remain</i>	[sala]	bhubha	<i>perish</i>	[buba]
fika	<i>arrive</i>	[fika]	bala	<i>count</i>	[ɓala]
hleka	<i>laugh</i>	[tɛka]	baba	<i>be acrid</i>	[ɓaɓa]
bonga	<i>praise</i>	[ɓɔŋga]	phola	<i>cool</i>	[pʰɔla]
buza	<i>ask</i>	[ɓuza]	bhala	<i>write</i>	[bala]
vala	<i>close</i>	[vala]	bila	<i>boil</i>	[ɓila]
pheka	<i>cook</i>	[pʰɛka]	beka	<i>put</i>	[ɓɛka]
bhonga	<i>roar</i>	[bɔŋga]	gula	<i>sick</i>	[gula]

- Speech was digitally recorded.
- Tags were inserted to identify the relevant segments.

### *Corpus*

Words containing the explosive [b] and implosive [ɓ] sounds (and other sounds) were embedded in a frame sentence which translates to *I say \_\_\_\_ now*. The subjects were asked to read the following sentences

Swati : *Ngitsi \_\_\_\_ nyalo* (where the blank contained words from Tables 1 and 2)

Xhosa: *Ndithi \_\_\_\_ ngoku* (where the blank contained words from Tables 1 and 2)

Zulu: *Ngithi \_\_\_\_ manje* (where the blank contained words from Tables 1 and 2)

### *Recording*

Three female mother-tongue speakers, of each of the Nguni languages, were used to record the corpus. All the subjects are students at the University of KwaZulu-Natal and reside in the eThekweni area. The Zulu subject is originally from the Hammarsdale area. The Xhosa subject is from the Grahamstown area and the Siswati subject originates from the Ngwavuma area. The speech was recorded directly onto a speech analysis toolkit, PRAAT<sup>7</sup> (version 5.1.04) at the Multimedia Learning Centre, University of KwaZulu-Natal. PRAAT is a software package used to record and analyse speech. PRAAT has different functions that enable the viewing of spectrograms and speech signals in time. The data was recorded in quiet conditions, using a ROSS RMA 200 microphone, at a sampling frequency of 22050Hz<sup>8</sup>. The words were recorded in a carrier phrase and were read in three phases. In phase one, the corpora shown in Tables 1 and 2 were read. In phase two, the corpora were read again. In phase three, the corpus shown in Table 1 was read. Each of the three recordings was made on separate days. This was followed by an analysis of spectrograms and waveforms.

## **2.2. Analysis of Data**

### *Burst Amplitude*

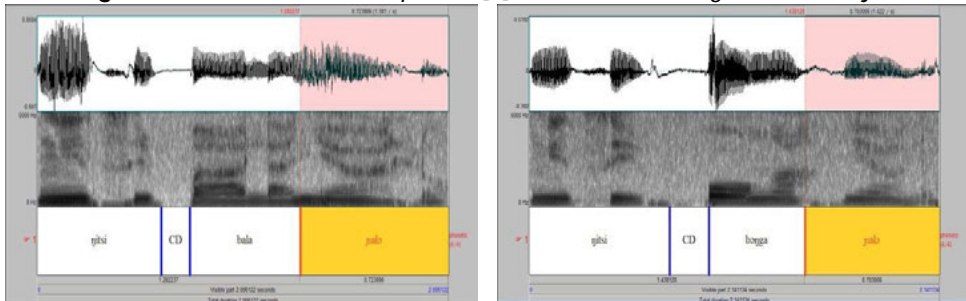
Under this parameter the following are dealt with:

- qualitative observations on the presence/absence of a burst or spike on the waveform and spectrogram
- provision of quantitative data on the intensity of the Zulu explosive [b] and implosive [ɓ]

A total of 90 spectrograms were analysed, 30 per speaker. For Swati, of the 15 spectrograms representing the three explosive [b] sounds; five depict a clear burst or spike confirming that these are explosive stop sounds. Two have a 'less prominent' burst and eight have no burst visible on the spectrogram. Of the 15 spectrograms representing the three implosive [ɓ] sounds; one depicts a clear spike

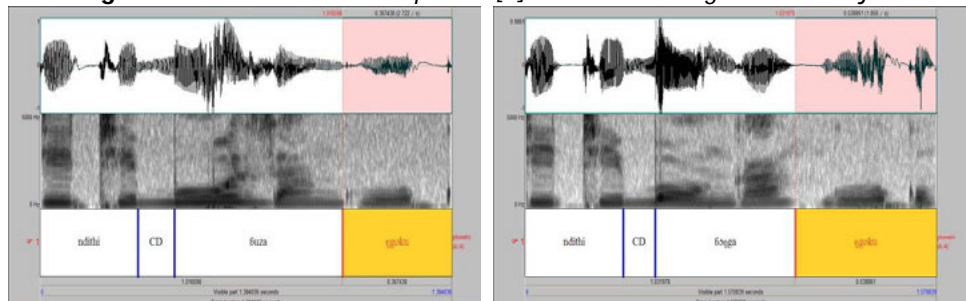
on the waveform, one has a ‘less prominent’ spike and 13 have no evidence of a spike.

**Figure 2:** *Absence of burst: explosive [b] in bhala and bhonga - Swati subject*



For Xhosa, all 15 spectrograms representing the three explosive [b] sounds depict a clear burst or spike confirming that these are explosive stop sounds. Of the 15 spectrograms representing the three implosive [ɓ] sounds; 10 depict a clear spike on the waveform, two have a ‘less prominent’ spike and three have no evidence of a spike.

**Figure 3:** *Presence of burst: implosive [ɓ] in buza and bonga– Xhosa subject*



For Zulu, of the 15 spectrograms representing the three explosive [b] sounds; 13 depict a clear burst or spike confirming that these are explosive stop sounds. One has a ‘less prominent’ burst and one has no burst visible on the spectrogram. Of the 15 spectrograms representing the three implosive [ɓ] sounds; two depict a clear spike on the waveform, three have a ‘less prominent’ spike and 10 have no evidence of a spike.

The parameter amplitude is closely related to that of Intensity. The latter, which is associated with the concept of loudness, is measured in decibels. The intensity values were obtained by clicking on the darkest portion of the spike or burst and using the Get Intensity function available in PRAAT. Shown in Table 3 are the average burst amplitude measurements for the explosive and implosive sounds. Table 3 indicates that for the Xhosa and Swati subjects, the implosive [ɓ] has an average lower amplitude burst compared with the explosive [b]. But the converse is true with the Zulu subject.

**Table 3:** *Average Intensity (decibels)*

	Swati	Xhosa	Zulu
bala	78.09	84.217	77.408
bhala	79.71	89.08	74.383
bhonga	80.585	89.017	68.012
bonga	78.84	83.515	75.587
buta/buza	78.186	83.369	71.566
bhubha	86.503	89.688	62.892

Shown in Table 4 are the overall average Intensity measurements for the implosive and explosive sounds for each of the Nguni languages. The results indicate that the Xhosa subject presents with a larger difference in the Intensity measurement between the implosive and explosive sounds. The difference is 5.56dB. The difference for Swati is 3.9dB.

**Table 4:** *Average Intensity (decibels)/Nguni language*

	Swati	Xhosa	Zulu
Implosive	78.37	83.7	74.85
Explosive	82.27	89.26	68.43

### 2.3. Closure Duration (CD)

CD values are quantified in Table 5. It indicates that, for all three Nguni languages, the CD for the implosive [ɓ] is consistently shorter than that of the explosive [b].

**Table 5:** *Average Closure Duration (sec)*

	Swati	Xhosa	Zulu
bala	0.19	0.114	0.094
bhala	0.163	0.175	0.111
bhonga	0.172	0.177	0.125
bonga	0.132	0.121	0.093
buta/buza	0.124	0.119	0.087
bhubha	0.152	0.172	0.112

Shown in Table 6 are the overall average CD measurements for the implosive and explosive sounds for each of the Nguni languages. The Swati and Zulu subjects present with an average difference of 0.01 and 0.03 sec between the CD of the implosive and explosive sound, respectively. The Xhosa subject presents with a longer average difference of 0.17 sec.

**Table 6:** *Average Closure Duration (sec)/Nguni language*

	Swati	Xhosa	Zulu
Implosive	0.15	0.35	0.09
Explosive	0.16	0.52	0.12

### 3. DISCUSSION

According to Ladefoged and Maddieson (1996), Johnson (1997) and Clements and Osu (2002), the implosive is characterised by a weak, less evident or absent burst on the spectrogram. The Naidoo (2010) experimental investigation for Zulu, corroborates that description. However, this comparative study<sup>9</sup> has produced slightly differing results. While the Zulu subject produced findings consistent with that of Naidoo (2010), namely 10 of the 15 spectrograms had no evidence of a burst for the implosive [ɓ] and 13 of the 15 spectrograms had a prominent burst for the explosive [b], the same was not applicable to Xhosa.

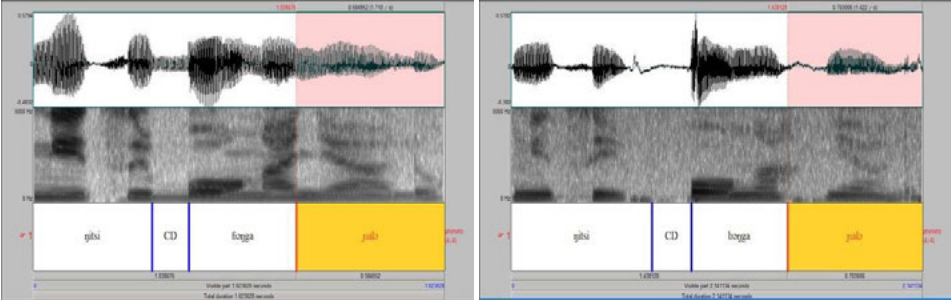
For the Xhosa subject, 10 of the 15 spectrograms had evidence of a burst for the implosive, while the Swati subject produced evidence of eight of the 15 spectrograms with no burst for the explosive [b]. Determining the presence of a burst on a spectrogram is a qualitative exercise and it can be argued that the determination is subjective. But as evidenced in Figure 2, there is no doubt that the Swati subject presents with the absence of a burst for the explosive [b] sound.

Similarly, the Xhosa subject presents with the presence of a burst for the implosive [ɓ] sound. This is shown in Figure 3.

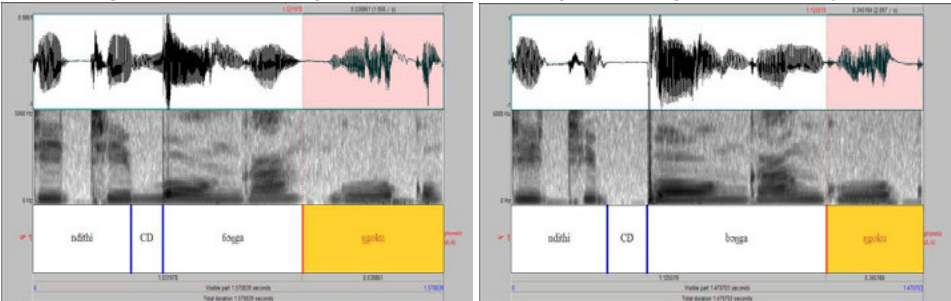
Any suspicion that the subjects have mispronounced the explosive and implosive sounds can be dismissed. The minimal pairs shown in Figures 4 and 5 indicate that the subjects are accurate in their pronunciation, producing spectrograms which contrast the implosive [ɓ] and explosive [b] sounds in terms of amplitude pattern and CD.



**Figure 4:** *Contrasting the minimal pair bonga vs bhonga – Swati subject*

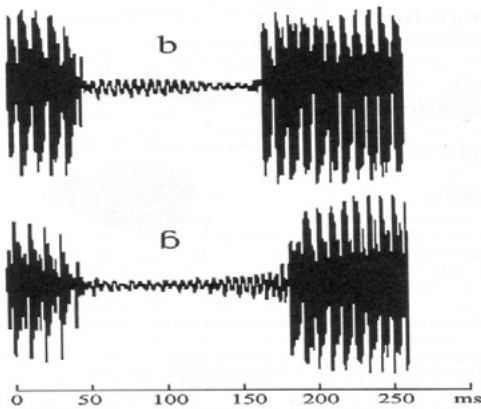


**Figure 5:** *Contrasting the minimal pair bonga vs bhonga – Xhosa subject*



The amplitude patterns are consistent with that of Ladefoged and Maddieson (1996:84) shown in Figure 6.

**Figure 6:** *Amplitude patterns*



This raises questions about the classification of the implosive and explosive in Xhosa and Swati respectively.

Jessen (2002:173) found that implosives have a lower burst amplitude value than explosives. While this experimental investigation found this true for Xhosa and Swati, the Zulu subject consistently had lower average intensity values for the explosive. This is reproduced in Table 7. This finding is inconsistent with that of Naidoo (2010). The other parameters, CD and burst pattern, for this subject, are consistent so it is possible that the intensity pattern is a personal idiosyncrasy.

**Table 7:** *Intensity values (decibels) – Zulu subject*

	Wordlist 1.1	Wordlist 2.1	Wordlist 1.2	Wordlist 2.2	Wordlist 1.3	Total	Average
bala	77.165	79.445	78.361	75.888	77.18	388.039	77.608
bhala	77.417	77.139	77.518	69.017	69.181	370.272	74.054
bhonga	65.138	69.351	76.666	64.479	64.426	340.06	68.012
bonga	77.93	77.595	84.555	73.62	64.234	377.934	75.587
buza	77.2	72.753	74.26	69.946	63.669	357.828	71.566
bhubha	67.948	62.009	64.195	64.017	56.292	314.461	62.892

#### 4. CONCLUDING REMARKS

The findings of the experimental investigation of Naidoo (2010) were the catalyst for this comparative study. The purpose of this experimental investigation was to determine whether or not the Nguni explosive [b] and implosive [ɓ] sounds present the same results in terms of the parameters burst amplitude and closure duration (CD). This comparative study has yielded observations which corroborate existing information:

- The implosive [ɓ] has a shorter CD value than the explosive [b]
- The implosive [ɓ] has a lower intensity value than the explosive [b]<sup>10</sup>
- The Swati and Zulu implosive [ɓ] is characterised by the absence of a burst on the spectrogram
- The Xhosa and Zulu explosive [b] is characterised by the presence of a burst on the spectrogram

However, two issues call for further investigation:

- Is the presence of a burst on the spectrogram of the implosive [ɓ] a widely occurring phenomenon in Xhosa?
- Is the absence of a burst on the spectrogram of the explosive [b] a widely occurring phenomenon in Swati?

## ENDNOTES

1. Crystal (1997:191) explains that an implosive sound is made using an airstream mechanism involving an inwards movement of air in the mouth (an ingressive airstream). A complete closure is made in the mouth, as with any plosive sound, but the air behind the closure is not compressed, ready for outward release; instead a downwards movement of the larynx takes place, and the air inside the mouth is accordingly rarefied. Upon release of the closure, air is then sucked into the mouth at the same time the glottis is released, allowing lung air to produce some vocal cord vibration.
2. Experimental work on implosives has been conducted by inter alia Lindau (1984), Pinkerton (1986), Nihalani (1986 & 1991), Wright and Shryock (1993), Demolin (1995), Best, McRoberts and Goodell (2001), Clements and Osu (2002), Cun Xi (n.d.) and Frazier (2009). Impressionistic observations and synchronic and diachronic studies of implosives emanate from inter alia Greenberg (1970), Kaye (1981), Goyvaerts (1988) and Kutsch Lojenga (1991).
3. The parameters amplitude pattern and Voice Onset Time were also investigated in Naidoo (2010) but the results were shown to be inconclusive. Therefore these parameters were excluded in this study.
4. In the next stage the corpus will be extended and the sample of informants widened to include males and speakers from more areas.
5. Traill, Khumalo and Fridjhon (1987), Giannini and Toscano (1988) and Best, McRoberts and Goodell (2001) have disputed the existence of the implosive [ɓ] in selected environments.
6. **Table 8:** Minimal pairs used in Doke's (1926) experimental study

Plosive [b]		Phonetic	Implosive [ɓ]		Phonetic
<i>bheka</i>	look	[beka]	<i>beka</i>	put	[ɓeka]
<i>bhabha</i>	entrap	[baba]	<i>baba</i>	be acrid	[ɓaba]
<i>bhiza</i>	have concern	[biza]	<i>biza</i>	call	[ɓiza]
<i>bhonga</i>	roar	[bɔŋga]	<i>bonga</i>	praise	[ɓɔŋga]
<i>bhuza</i>	buzz	[buza]	<i>buza</i>	ask	[ɓuza]

7. <http://www.praat.org> Cun Xi (n.d.) and Frazier (2009) have used PRAAT for their recording and analysis. Hence my use of PRAAT is consistent with other researchers investigating the implosive.
8. Other researchers have used this sampling rate.
9. Readers who wish to view all the spectrograms may email the author.
10. The Zulu subject presents a personal idiosyncrasy. The findings here are not consistent with other Zulu subjects.

## REFERENCES

- Best, C.T., G.W. McRoberts & E. Goodell. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America* 109:775-794.
- Boersma, P. & D. Weenink. 1992. *PRAAT doing phonetics by computer* Version 5.1.04. <http://www.praat.org/> Accessed: 20-01-2011.
- Clements, G.N. & S. Osu. 2002. Explosives, implosives and nonexplosives: the linguistic function of air pressure differences in stops, in *Laboratory Phonology VII*, edited by C. Gussenhoven & N. Warner. Berlin: Mouton de Gruyter.299-350.
- Clements, G.N. & A. Rialland. 2005. Africa as a phonological area. <http://nickclements.free.fr/publications.html> Accessed: 31-03-2009.
- Crystal, D. 1997. *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell Publishers.
- Cun Xi, n.d. The phonetic characteristics of implosives in Tow Chinese dialects. [http://seneca.uab.es/filologia\\_catalana/papi/files/cun-xi.pdf](http://seneca.uab.es/filologia_catalana/papi/files/cun-xi.pdf) Accessed: 06-04-2009.
- Demolin, D. 1995. The phonetics and phonology of glottalized consonants in Lendu, in *Phonology and Phonetic Evidence Papers in Laboratory Phonology IV*, edited by B. Connell & A. Arvaniti. Cambridge University Press.368-385.
- Doke, C.M. 1926. *The Phonetics of the Zulu Language*. Johannesburg: University of Witwatersrand Press.
- Doke, C.M., D.M. Malcolm, J.M.A. Sikakana & B.W. Vilakazi. 1990. *English-Zulu Zulu-English Dictionary*. Johannesburg: Witwatersrand University Press.
- Finlayson, R., J. Jones, K. Podile & J.W. Snyman. 1993. *An Introduction to Xhosa Phonetics*. Cape Town: Marius Lubbe Publishers.
- Fischer, A., E. Weiss, S. Tshabe & E. Mdala. 1985. *English Xhosa Dictionary*. Cape Town: Oxford University Press.
- Frazier, M. 2009. Tonal dialects and consonant-pitch interaction in Yucatec Maya. <http://www.unc.edu/~melfraz/ling/Frazier-TonalDialects.pdf> Accessed: 06-04-2009.
- Giannini, A.M.P. & M. Toscano. 1988. Some remarks on Zulu stops. *Afrikanistische Arbeitspapiere* 13:95-116.
- Goyvaerts, D. 1988. Glottalized consonants: a new dimension. *Belgian Journal of Linguistics* 3:97-102.
- Greenberg, J.H. 1970. Some generalizations concerning glottalic consonants, especially implosives. *International Journal of American Linguistics* 36(2):123-145.

- Jessen, M. 2002. An acoustic study of contrasting plosives and click accompaniments. *Phonetica* 59:150-179.
- Johnson, K. 1997. *Acoustic and Auditory Phonetics*. Cambridge: Blackwell Publishers.
- Kaye, J.D. 1981. Implosives as liquids. *Studies in African Linguistics*, Supplement 8:78-81.
- Kutsch Lojenga, C. 1991. Lendu: a new perspective on implosive and glottalized consonants. *Afrika und Übersee* 74(1):77-86.
- Ladefoged, P. & I. Maddieson. 1996. *The Sounds of the World's Languages*. Oxford: Blackwell Publishers.
- Lindau, M. 1984. Phonetic differences in glottalic consonants. *Journal of Phonetics* 54:147-155.
- Naidoo, S. 2010. A re-evaluation of the Zulu implosive [ɓ]. *South African Journal of African Languages* 30(1):1-10.
- Nihalani, P. 1986. Phonetic implementation of implosives. *Language and Speech* 29:253-262.
- Nihalani, P. 1991. A re-evaluation of implosives in Sindhi. *UCLA Working Papers in Phonetics* 80:1-5.
- Pinkerton, S. 1986. Quichean (Mayan) glottalized and nonglottalized stops: a phonetic study in *Experimental Phonology*, edited by J.J. Ohala & J.J. Jaeger. Orlando: Academic Press. 125-139.
- Rycroft, D.K. 1981. *Concise SiSwati Dictionary*. Pretoria: JL van Schaik Publishers.
- Taljaard, P.C. & J.W. Snyman. 1989a. *An Introduction to SiSwati Phonetics*. Cape Town: Marius Lubbe Publishers.
- Taljaard, P.C. & J.W. Snyman. 1989b. *An Introduction to Zulu Phonetics*. Cape Town: Marius Lubbe Publishers.
- Traill, A., J.S.M. Khumalo. & P. Fridjhon. 1987. Depressing facts about Zulu. *African Studies* 46(2):255-274.
- Wright, R. & A. Shryock. 1993. The effects of implosives on pitch in SiSwati. *Journal of the International Phonetic Association* 23(1):16-23.

# CHAPTER 3

## ON THE PHONETIC AND THE PHONOLOGICAL STUDIES OF SOUTHERN AFRICAN LANGUAGES

John Lubinda

*Department of French, University of Botswana, Gaborone, Botswana*  
lubindaj@mopipi.ub.bw

### 1. INTRODUCTION

This chapter reflects on the phonetic and phonological descriptions of Southern African languages. It is common knowledge that one of the priority tasks for any linguist conducting fieldwork in sub-Saharan Africa is to collect speech data on a specific language. These data will further be exploited to analyse and describe the language characteristics using appropriate technical devices and analytical procedures. The description will need to apply objective criteria within the framework of a fitting, recognised and well-established theoretical model. This is one of the fertile grounds for research today in descriptive African linguistics.

The Southern African sub-region, however, still has many insufficiently documented indigenous languages, especially the so-called *minority languages*. Some of these languages have not even been codified yet or standardised. Any attempt to promote them, in a meaningful way, must start with analysing and describing them at the phonetic and phonological levels. This involves, among other things:

- (i) identifying functional contrasting sound units in the languages being described;
- (ii) analysing and describing the identified linguistic sounds, in terms of their distinctive articulatory and acoustic characteristics and other relevant phonetic features, including suprasegmental features;
- (iii) determining the pattern of their organisation within the languages being investigated, forming hypotheses about the way the sound systems of these languages work and testing these hypotheses with scientific rigour;
- (iv) formulating phonological rules to account for the internal workings of the sound system of each language being studied.

Professional linguists readily recognise that phonetics, along with the cognate subject of phonology, is the logical and most obvious starting point in the study of

any given language. In fact, as Guma (1971:8) aptly points out, “*the sounds of a language serve as the building blocks or bricks out of which our speech is put together*”. A similar viewpoint, with regard to the study of speech sounds, is expressed more emphatically by Brosnahan and Malmberg (1970:8) when they observe that “*phonetics is a basic branch of the science of linguistics: neither linguistic theory nor linguistic practice can do without phonetics, and no language description is complete without phonetics*”. In this statement the term ‘phonetics’ is to be understood in the broader sense, to cover both the traditional domains of phonetics (articulatory, acoustic and auditory) and what Malmberg (1963:91) calls “*functional phonetics*”, better known today as ‘*phonology*’ or formerly ‘*phonemics /phonematics*’.

As Westermann and Ward (1933:6) correctly pointed out, several decades ago, “*phonetics is essentially a practical science*”. Indeed, there are many practical applications of knowledge and insights gained in phonetics and phonology to the solution of some practical problems of language development and language teaching, notably in the areas of pronunciation, phoniatrics and logopedics (Brosnahan & Malmberg, 1970:9-10; O’Connor, 1973:276-282). Speech research is therefore more than just a mere theoretical and elitist discipline, an academic exercise, so to speak, and the exclusive preserve of speech scientists, supposedly engaged in their “*ivory tower*” pursuit of basic research, delving in so-called “*phonetic minutiae*”. On the contrary, it has utilitarian spinoffs in the real world of professional work, such as in speech therapy (language re-education and training), audiology, speech technology, telecommunications engineering, speech-training for elocutionists, automatic speech recognition and in speech synthesis (Catford, 1988:1-2). It is certainly not a casual pastime-like activity, as glamorised by Professor Henry Higgins, the fictional character passing for an eccentric phonetician, portrayed in George Bernard Shaw’s famous play, *Pygmalion*, and later adapted as a musical comedy, in the well-known classic movie *My fair lady*.

It is a fact that phonetics/phonology research is one of the most neglected parts of language research in Southern Africa. Programmes of few recent language conferences in Southern Africa (ALASA 2010 and 2011, LASU 2011 and AFRILEX 2010 and 2011)<sup>1</sup> show very few papers related to phonetics and phonology. It is equally true that very few universities in the SADC region have phonetics laboratories or research units devoted to phonetics and phonology. Thus, the intended focus of this chapter is to plead the case for a more scientific approach to the study of the phonetic/phonological aspect of individual Bantu languages.

## **2. SETTING OUT OBJECTIVES AND PRIORITIES**

For anyone contemplating to undertaking research in any Bantu language, at the phonetic and phonological level, a preliminary first step is to set out clearly the aims and objectives of the proposed study, as well as establishing clearly a hierarchy of priorities. From the outset, it is important to take cognisance of what has been done previously on the language (i.e. a historiography of linguistic research in the language by predecessors) and determine what still remains to be done. Some of these languages have not even been committed to print or codified yet, others have been insufficiently or merely superficially documented whereas others, albeit a comparatively small minority, have been the subject of much research, in relatively recent years. Among those that have been studied and analysed at the phonetic/phonological level, it will be discovered, however, that some have been inaccurately described or characterised, in certain respects, albeit unwittingly. There is therefore some considerable remedial work to be done here.

The following set of basic pertinent questions ought to be kept in the forefront of one's mind when undertaking the study of a given Bantu language, or of any other language for that matter, at the phonetic and/or phonological level:

- (i) Does the language have a written form? In other words, has it been reduced to writing by transcribers?
- (ii) Is the language codified and standardised, in terms of orthography and grammar? What is the extent of dialect variation observable in the language?
- (iii) Has the language ever been documented (described and analysed) at the phonetic/phonological level? If so, how reliable or valid are these earlier linguistic accounts, in terms of methodologies and procedures followed, in light of current trends and knowledge?
- (iv) What are the main research priorities and focus areas in the language and contemporary issues at the time of undertaking the research, and to which end?
- (v) Which research methods and techniques are best suited for recording, storing and analysing phonetic data, and which descriptive models does one wish to adopt, to ensure theoretical adequacy?

### **2.1. Research Priorities and Focus Areas in Phonetic/Phonological Research**

Research priorities in well-researched languages of wide communication and distribution, such as English, French or Portuguese, to cite only the official working languages of the SADC Secretariat, are not necessarily the same as those in a



language that is still inadequately researched and perhaps one which is even not yet fully codified. A linguistic community, whose language is not yet written, for instance, would appreciate more research efforts devoted to facilitating, in the first instance, the codification of the home language and giving it a practical and user-friendly orthography for use by native speakers and by learners of the language. This is, arguably, one of the areas of practical application where insights gained from phonetics and phonology may be of great value. It is axiomatic that knowledge of the sound system of a given language is also important in the teaching of that language.

The other key priority area in phonetic research on Southern African languages (including Khoisan) is dialectology (also variously called linguistic geography, dialect geography or dialectography). Malmberg (1963:107) points out that phonetics is indispensable in the study of language variation (regional, social or stylistic varieties of the same language) and sound change in the evolution of a given language. In descriptive or synchronic linguistics, phonetics is also important in the norm planning of a language i.e. defining the standard form of the language, having surveyed the entire dialect field. Any meaningful work on a little-known language that is still in need of standardisation requires insights deriving from phonetics and phonology.

The other priority area for phonological research in these languages is to study their sound patterns, using appropriate procedures and theoretical perspectives. The study of the sound pattern implies, among other things, identifying and examining:

- (i) the set of phonemes occurring in a given language and their distinctive features (segmental phonology)
- (ii) the rules governing permissible and proscribed sound sequences in the speech chain of a given language
- (iii) the rules pertaining to the processes of changing, adding or deleting of sound segments in connected speech (such natural processes like strengthening, palatalisation, velarisation, etc.)

Moreover, at the suprasegmental level, the researcher should study systematically the prosodic elements manifest in the language. These are vocal features such as tone, intonation and stress which are superimposed on the segmental units in the flow of speech. This is the realm of non-segmental phonology.

For (i), (ii) and (iii) above, the researcher must proceed to a comprehensive phonemic analysis of the language, involving the following major phases:

- (i) Establishing a complete inventory of segmental phonemes by means of the commutation test on minimal pairs, taking into account distributional constraints

In the process, the researcher will distinguish between phonemes and variants of the same phoneme, in complementary distribution or in free variation. Initially it involves the task of painstakingly and meticulously transcribing all speech sounds attested and recorded in the corpus of the language under study, as they are differentiated by the trained ear. Batibo (2000:182) points out that:

The basic task of the phonologist is to discover and then describe the phonological system of the language he or she is working on. Once the extensive fieldwork has been completed and all the speech sounds (phones) of the language of study have been collected and recorded (on tape or by some other means) the important question is, which of the phones are phonemes?"

The following tasks will have to be carried out in the process:

- (ii) Classification of identified phonemes and accounting for allophonic variation
- (iii) Formulating phonological rules of two types, namely allophonic rules and morphophonemic (or morphophonological) rules
- (iv) Making a phonotactic analysis, i.e. a systematic account of how sounds combine in sequence in the language to form words and utterances, especially an account of the syllable structure of the language
- (v) Carrying out a comprehensive prosodic analysis of the language, based on empirical data, following an appropriate methodological protocol, investigating such dynamic features as tonal registers and movements, vowel length, articulatory force (stress) and intonation
- (vi) Investigating coarticulatory processes in connected speech: assimilation, vowel coalescence and fusion, prenasalisation in certain consonantal compounds, etc.

The rather obvious logical starting point for any researcher engaged in phonetic fieldwork is, of course, to record the speech data to be described and analysed, using appropriate recording tools and techniques. This is to be done, of course, with the active involvement of consenting native speakers of the language, as producers of samples of language material and as informants. For the results of the study to be deemed valid, proper elicitation techniques ought to be used in collecting data, and in some cases, controlled production conditions must be created and observed. Conditions of experimentation, later on, during the data analysis phase, must be made explicit and observed too, just like in any other

experiment of any scientific discipline. Moreover, there ought to be some clarity on the type of conceptual framework being applied at the phonological level.

Phonology is a rapidly growing sub-field of linguistics, open to new theories and approaches and it is constantly adding new terminology to its stock of metalanguage. Compared to other aspects of linguistic theory, phonological theory has indeed evolved very rapidly and has undergone radical changes within the past few decades. Ever since the publication of Chomsky and Halle's (1968) *The Sound Pattern of English* (SPE), that inaugurated the era of Generative Phonology (GP), various strands and offshoots of the SPE model of standard GP have emerged, such as natural generative phonology, autosegmental phonology, natural phonology, metrical phonology, CV phonology, optimality theory, etc. Some of these models of phonological description and analysis, such as autosegmental phonology, can be conveniently adapted to a phonological study of a given Bantu language. See, for instance, Miti (1988).

## 2.2. Importance of Empirical Research

Speech science being a full-fledged science in its own right (albeit a human and social science) its methods and heuristic procedures must, of necessity, conform to the principles and canons of scientific enquiry: Explicitness, systematicness and objectivity (see Crystal 1971:78), based on careful observation and systematic analysis of acquired data, measurement and testing for verification of initial hypotheses.

In a quantitative perspective, a statistical approach may be required to support and validate the conclusions of the findings. Researchers working on Bantu languages must therefore begin to adopt the scientific and systematic approach when describing and analysing the grammatical structure and sound pattern of any language they are investigating.

A corpus-based approach and the use of appropriate human language technologies are crucially important here. Impressionistic and subjective (not to say "*pedestrian*") methods of determining the nature and quality of sound units, using only the listener's ear (be it "*a trained ear*") and direct articulatory observation by his/her naked eye, must now be supplemented with objective instrumental analysis of these speech units and events in the phonetics laboratory. The perceptual capacity of the human ear and that of direct articulatory or kinaesthetic observation, however useful, have nevertheless some obvious limitations.

In articulatory phonetics (termed "*physiological phonetics*" by some authors), the scientifically acknowledged methods of anatomy and physiology (of the human vocal apparatus) need to be applied to enhance objectivity and descriptive precision. Thanks to the use of such specialised laboratory techniques as cine-

radiography (with synchronised sound track), electromyography, miringography, laryngoscopy and palatography, the physiology and aerodynamics of speech production can now be observed and analysed with greater precision and accuracy than was previously possible. Indeed, in certain respects, the study of speech sounds has attained a level of sophistication and objectivity comparable to that of the so-called exact sciences, like physics, chemistry and mathematics.

Similarly, in acoustic phonetics, succinctly described by Fowler (1974:162) as the "speech-oriented sub-division of physics", appropriate methods of physics ought to be harnessed. Since the close of the Second World War, acoustic phonetics has witnessed rapid progress, owing largely to the tremendous advances made in electro-acoustics and in computer technology. Various sorts of high-tech instrumental devices and accompanying software have been designed over the years for the analysis of speech sounds in this aspect of phonetics. Detailed descriptions of these technical instruments and of the way they are used in the phonetics laboratory can be found on internet and in a number of works in phonetic literature such as Mettas (1971), Emerit (1977), Fujimura and Erickson (1997), Hirose (1997), Ladefoged (1997 & 2003) and Stone (1997).

The major structural properties of the sound wave, such as frequency and amplitude of vibration, can now be directly and easily observed, and analysed fairly accurately, with the aid of a number of analytical laboratory instruments, the most emblematic of which is the sound spectrograph. The information obtained by spectrographic analysis of the sound signal can be correlated with (or used to supplement) that yielded by the articulatory study so as to arrive at a more complete picture of the nature and characteristics of the speech sounds being investigated. It is also possible now to describe speech sounds objectively either from the articulatory point of view (i.e. in terms of the mode of their production by the vocal organs) or from the acoustic perspective, in terms of disturbances or turbulences caused in the air, in the form of sound waves transmitted from the mouth of the speaker to the listening ear(s) of the hearer. It is also possible to study speech phenomena from the auditory perspective, studying such parameters as loudness which correlates with intensity and amplitude length (see Stevens 1997 for details). These three view points are mutually complementary.

For illustrative purposes, the following descriptive template was adopted by Lubinda (1987:275-276) in a preliminary articulatory study of Silozi consonants. Silozi is a Bantu language spoken principally in South-western Zambia, Eastern Caprivi of Namibia and in a small community of North-western Botswana. It is classified as K21 by Guthrie (1971).

Radiographic tracings of a dynamic film strip were carefully made by the researcher. It was directly on these tracings that the following measurements were made:

- (i) The horizontal distance between the upper incisor and the upper lip (projection of the upper lip)
- (ii) The horizontal distance between the lower incisor and the lower lip (projection of the lower lip)
- (iii) The vertical distance between the lower lip and the upper lip
- (iv) The vertical distance between the lower incisor and the upper incisor
- (v) The distance between the tongue and various parts of the roof of the mouth: within the dental zone, within the pre-palatal zone, within the palatal or velar zone,
- (vi) The distance between the velum and the pharyngeal lining, at the point of maximum narrowing of the air passage
- (vii) Pharynx constriction diameter: the distance between the tongue root and the pharyngeal lining at two specific points.
- (viii) Position of the hyoid bone, on two dimensional planes: the horizontal and vertical axes

It will be noted that some of these measurements apply only to certain consonantal articulations. For instance, measurement in (vi) applies only to nasals, whereas the measurement concerning lip protrusion applies only to labial fricatives like [ɸ] and [w].

The X-ray motion picture film had been shot at the rate of 50 images per second. Thus each image represented 2 centiseconds. It was therefore possible to measure the duration of articulations or of phases of articulations. This physiological technique was combined with the results of palatographic and mionographic analyses. These two techniques enabled the researcher to obtain data on laryngeal activity, nasalisation, segment length and air pressure variations. With a combination of these three techniques, it was possible to describe particular phones such as the consonant [ɸ] as a pre-palatal voiceless fricative. This contrasts with a description that is merely impressionistic. Fortune (2001), who uses the latter approach, describes the Silozi sibilant [ɸ] as palatal, perhaps because in many other languages it is described as such. However, observable and verifiable experimental data consistently show that it is *pre-palatal* rather than *palatal*. This is just one example among many others.

Previous phonetic/phonological descriptions and analyses ought to be reviewed in light of the results obtained through instrumental analyses, to ascertain their accuracy. While acknowledging, quite correctly, that we owe an enormous debt of gratitude to pioneers in the study of Bantu languages we have, nevertheless, an obligation to check the accuracy of their descriptions and analyses, in light of the insights we have gained through instrumental analysis of speech. Computer

technology opens up new avenues for better and finer analyses, especially in the areas of acoustics and auditory phonetics. As Edwards and Shriberg (1983:13) point out: *"Present-day computers can be specially programmed to complete with a high degree of reliability a variety of acoustic analyses of speech events"*.

More recently, Zsiga (2006:35) explains that:

In the first part of the 21st century speech analysis is done by computer. Microphones still convert the vibration of the membrane into variations in electrical current. But then computers convert a continuously varying sound wave into a series of numbers: this is called an analogue-to-digital (A to D) conversion. (Working in the opposite direction getting a computer: digital video-disk or digital audio player to make sound is D to A conversion).

Once represented and stored in a digital format, sound files can be mathematically analysed to separate out the different frequencies, and the results of analysis displayed on screen in various formants.

For computer-assisted acoustic and perceptual studies one needs at least a good computer-based speech workstation with appropriate software. Improved instrumentation and techniques of analysis have made it possible to examine, in a quantitative way, the articulatory motions and configurations, the mechanics and aerodynamics of the respiratory system, the activity of the muscles that control the structures utilised in speech production. Indeed, acoustic phonetics, in particular, has made significant progress during the past half century thanks, in part, to the improvements and advances made in electronic and computational technology. A whole range of laboratory instruments and apparatuses are currently able to assist the phonetician in his/her investigation of sound phenomena.

Brosnahan and Malmberg (1970:6) point out that

Instrumental methods deriving from physiology and physics were introduced into phonetics in the second half of the last century in order to supplement and indeed to rectify the impressions deriving from the human senses, especially the auditory impressions, since these were held to be affected by the limitations of the perceptual mechanism, and in general, subjective.

Ever since the development of this approach to the study of speech sounds, phonetics has been closely associated, in the mind of some lay people, not well-acquainted with the subject, with expensively equipped (some would even say "extravagantly equipped") laboratories in which speech is recorded, transcribed in esoteric-looking script and minutely dissected. However, phonetics is not merely a matter of toying with electronic and computer gadgets, digital video cameras and speech synthesisers – a simple fascination with technical devices. The use of laboratory instruments and techniques is only a means to an end, rather than an end in itself.

As Brosnahan and Malmberg (1970:7) also add:

Phonetics ... is not primarily the study of processes in the vocal apparatus or in the hearing apparatus or of sound waves as simple physical events, but the study of these phenomena as limited, modified or distorted by the ways in which the human brain perceives and controls them.

### 3. METHODS OF INVESTIGATION AND DEVICES IN EXPERIMENTAL PHONETICS

There are several methods and tools used in the analysis of speech sounds and the study of processes of their production. This section does not attempt to present an exhaustive list or description of all those methods. It is merely a very brief sampling, for illustrative purposes. These may be conveniently classified into two broad categories, namely methods and devices of acoustic analysis of the sound signal or sound wave such as:

- The **sound spectrograph** for the visual display of speech spectra. It is a useful instrument for sound print analysis. Three dimensions of sounds may be represented on the spectrogram: Duration, frequency and intensity.
- The **oscilloscope** for the visual display of the frequency and amplitude of the wave form. The visual image ("visible speech", to use the old-fashioned terminology of the Bell Telephone Laboratories) is termed "oscillogramme".

There are also techniques and devices for articulatory study of speech sounds like the following three:

- **Palatography**: A technique that provides information on points of contact between the tongue and different parts of the roof of the mouth, such as the upper teeth, the alveolar ridge or the hard palate. It is a 19th century technique which, in its improved form, is still much used for determining the point of closure in the oral cavity for lingual articulations, such as stop consonants, nasals and laterals. Ladefoged (2003) describes different systems of palatography, from basic palatography to the more elaborate palatography, such as dynamic electropalatography.
- **Cineradiography** (dynamic X-ray photography) which shows the positions and movements of the active articulators during speech production. Radiographic tracings can be made of the target positions for the sound segments being studied. A great deal of precise information concerning, for instance, mandible and tongue movements, place of articulation, raising and lowering of the velum, lip posture and lip closure, etc. is obtained from a visual display of series of radiographic images.
- **Mingography**: A technique in the category of oscilloscopy for the recording and display of the sound wave and yielding information on four line tracings, using the principle of the old-fashioned kymograph. An accurate interpretation of

information displayed on the four line tracings enables the speech analyst to determine the mode of articulation of a given sound segment, its duration of emission, the nature of laryngeal activity during its production, vocalic opening or closure, as well as the degree of expiratory force when the sound is produced. For stop consonants there is useful information on the duration of the hold phase and on the nature of the explosion in the release phase. Mettas (1971) considers the mingograph as a very useful instrument in phonetic research, not only for sound segments but also for prosodic analysis, when coupled with other technical devices used in phonetic research.

There are other technical devices for acoustic analysis and for use in articulatory studies that limitations of space do not allow to list or describe here. There are also specialised instruments for measuring airflow and air pressure variations in aerodynamic studies. Detailed descriptions of techniques and technical devices of the phonetics laboratory, for speech analysis and for the study of the anatomy and physiology of speech production are found in, among others, Ashby and Maidment (2005), Johnson (1997) and Ladefoged (2003). Language researchers on Bantu languages in Southern Africa cannot therefore continue to rely solely on subjective methods of direct observation and sensory perception when tools are actually available to conduct scientific investigations of the sound systems of these languages. At the moment, the sounds and prosodic elements occurring in these languages can be described precisely and objectively, based on the results of the analysis of empirical data in the phonetics laboratory.

#### **4. CHALLENGES AND STRATEGIES**

The major challenge facing the phonetician working on any of the phonetically undocumented (or insufficiently documented) Bantu languages is to adopt a method that will permit him or her to describe accurately and precisely the sound features of the language being investigated. To achieve this, s/he needs to have at least the most basic infrastructural and technical facilities of the modern phonetics laboratory, particularly equipment and software enabling him or her to conduct relevant research in descriptive and/or experimental phonetics. Good sound-recording and sound-reproducing equipment is of utmost importance. Ladefoged (2003:183) suggests a list of “basic instrumentation for phonetic data collection and analysis” that is worth to be consulted.

To this list the researcher(s) will, of course, add such pieces of valuable technical equipment as the sound spectrograph and the mingograph. Every department of linguistics in which there are academics engaged in phonetic research ought to have a basic general purpose phonetics laboratory with at least the basic language technologies. It should be equipped according to the particular needs of the



department, in terms of identified research priorities. Moreover, it is necessary to be well-trained and skilful in the use of research techniques of the phonetics laboratory. This involves, among other things, knowing how to:

- (i) select speakers to be recorded
- (ii) make recordings, including digital recordings
- (iii) listen to recordings
- (iv) analyse and interpret recorded data
- (v) report and disseminate the findings

Like for any other science laboratory, there is a need to have a lab assistant or technician for the maintenance and repair of equipment, and for technical advice whenever necessary, concerning the operation and optimum use of the equipment.

What has been said above, concerning training needs, in the area of phonetics, is *mutatis mutandis*, equally true in the area of phonological theory. The researcher should be well acquainted with the current trends in phonological theory and be able to determine which phonological models of description to adopt for the description and analysis of the sound pattern of the Bantu language s/he has chosen to investigate.

## 5. CONCLUSION

The present chapter began with the observation that relatively few phonetic studies have been undertaken on Bantu languages spoken in Southern Africa. It has emphasised the importance of linguistic studies at the phonetic and phonological levels. Moreover, it has attempted to identify some of the current priority and focus areas in phonetic and phonological research in these languages. It has suggested some of the tasks to be carried out by researchers in this area. It briefly indicates some of the methods and techniques employed in phonetic research and ends with a few proposals on how to advance research in phonetics and phonology in Bantu linguistics within the sub region of Southern Africa, using human language technologies.

## ENDNOTE

1. The following abbreviations are used in this chapter: *ALASA* for the African Languages Association of Southern Africa, *LASU* for the Linguistics Association of SADC Universities, *AFRILEX* for the African Association for Lexicography and *SADC* for the Southern African Development Community.

## REFERENCES

- Ashby, M. & J. Maidment. 2005. *Introducing Phonetic Science*. Cambridge: Cambridge University Press.
- Batibo, H. M. 2000. System in the sounds of Africa. *African Voices : An Introduction to the Languages and Linguistics of Africa*, edited by V. Webb & J. Kembo-Sure. Cape Town: Oxford University Press Southern Africa (Pty) Ltd.160-196.
- Brosnahan, L.F. & B. Malmberg. 1970. *Introduction to Phonetics*. London: Cambridge University Press.
- Catford, J.C. 1988. *A Practical Introduction to Phonetics*. Oxford: Clarendon Press.
- Chomsky, N. & M. Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.
- Crystal, D. 1971. *Linguistics*. Harmondsworth: Penguin Books.
- Edwards, M. L. & L.D. Shriberg. 1983. *Phonology: Applications in Communicative Disorders*. California: College-Hill Press, Inc.
- Emerit, E. 1977. *Cours de Phonétique Acoustique*. Alger: Société Nationale d'Édition et de Diffusion.
- Fortune, G. 2001. *An outline of Silozi Grammar*. Lusaka: Bookworld Publishers.
- Fowler, R. 1974. *Understanding Language: An Introduction to Linguistics*. London: Routledge & Kegan Paul Ltd.
- Fujimura, O. & D. Erickson. 1997. Acoustic Phonetics. *The Handbook of Phonetic Sciences*, edited by Hardcastle, W.J. & J. Laver. Oxford: Blackwell Publishing.65-115.
- Guma, S. M. 1971. *An outline Structure of Southern Sotho*. Pietermaritzburg: Shuter and Shooter.
- Guthrie, M. 1971. *Comparative Bantu* (4 vols.). Farnborough: Cregg International Publishers Ltd.
- Hirose, H. 1997. Investigating the physiology of laryngeal structures. *The Handbook of Phonetic Sciences*, edited by Hardcastle, W.J. & J. Laver. Oxford: Blackwell Publishing.116-136.
- Johnson, K. 1997. *Acoustic and Auditory Phonetics*. Oxford: Blackwell Publishers.
- Ladefoged, P. 1997. Instrumental techniques for linguistic Phonetic fieldwork. *The Handbook of Phonetic Sciences*, edited by Hardcastle, J.W. & J. Laver. Oxford: Blackwell Publishing.137-166.

- Ladefoged, P. 2003. *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Victoria: Blackwell Publishing.
- Lubinda, J. 1987. *Étude Articulaire des Consonnes du Lozi: Analyses mingographique, Radiocinématographique et palatographique*. Tomes I & II, PhD Thesis, Institut de Phonétique, Université des Sciences Humaines de Strasbourg II.
- Malmberg, B.F. 1963. *Phonetics*. New York: Dover Publications, Inc.
- Mettas, O. 1971. *Les Techniques de la Phonétique Instrumentale et l'Intonation*. Bruxelles: Presses Universitaires de Bruxelles.
- Miti, L. 1988. *Tonal Variation in Zambian Chinyanja varieties: An Autosegmental Analysis*. PhD Thesis, University of London.
- O'Connor, J. D. 1978. *Phonetics*. Harmondsworth: Penguin Books.
- Shearer, W. 1997. Experimental design and statistics in Speech Science. *The Handbook of Phonetic Sciences*, edited by Hardcastle, W.J. & J. Laver. Oxford: Blackwell Publishing.167-187.
- Stevens, K.N. 1997. Articulatory – Acoustic – Auditory Relationships. *The Handbook of Phonetic Sciences*, edited by Hardcastle, W.J. & J. Laver. Oxford: Blackwell Publishing.462-506.
- Stone, M. 1997. Laboratory Techniques for Investigating Speech Articulation. *The Handbook of Phonetic Sciences*, edited by Hardcastle, W.J. & J. Laver. Oxford: Blackwell Publishing.11-32.
- Westermann, D. & I.C. Ward. 1933. *Practical Phonetics for Students of African Languages*. New York: Oxford University Press. Publication for the International African Institute.
- Zsiga, E. 2006. The sounds of Language. *An Introduction to Language and Linguistics*, edited by Fosold, R.W. & J. Connor-Linton. Cambridge: Cambridge University Press.13-53.

## CHAPTER 4

# ON THE MANY FACES OF TONE IN SOUTHERN SOTHO: A CASE STUDY

Daan Wissing

*Centre for Text Technology, North-West University, Potchefstroom, South Africa*  
daan.wissing@nwu.ac.za

### 1. INTRODUCTION

This study is an acoustic description and assessment of aspects of tone in Southern Sotho, a Southern Bantu language belonging to the Sotho-Tswana group. Although various acoustic studies exist on Southern Sotho (cf. Roux 1979; Wissing 2005, 2010a and 2010b; and Barnard and Wissing 2008 to name a few), there is no existing acoustic study available on this specific matter. Roux (1983), in an unpublished report suggested the Perceptual Confusion Hypothesis forms an exception to this. Though he focused on perceptual issues, he also suggested a production survey, such as the current one. In essence this hypothesis implies an interaction between tone and vowel quality. For example, high tone specifically may influence the listener to perceive a higher articulated vowel than actual, and vice versa, a higher produced vowel may lead to perceiving a higher than actual tone. Tucker's (1929:23) remark "Sometimes a high tone on the syllable containing *e* will also produce this effect, e.g. *ke tau!* (that's a lion!)," relates to the latter situation. It is a secondary aim of the present chapter to prepare the ground for a perception study in which Roux's Perceptual Confusion Hypothesis can be tested.

### 2. BASIC OUTLINE OF TONE AND VOWEL QUALITY IN SESOTHO

The present presentation does, of course, not intend any elaborate exposition of the very intricate sound system of Sesotho. Good descriptions can be found in the work of Tucker (1929); Letele (1955); Doke and Mofokeng (1957); Guma (1971); Roux (1983); Khabanyane (1991); Selebeleng (1997); Krüger and Snyman (s.a.), and Cole (1955) for Tswana, and Poulos and Louwrens (1994) for Northern Sotho. My contribution is confined to a very small segment of Southern Sotho, as spoken in the Eastern Free State, South Africa. Tucker's (1929) reference to "[t]he wonderful similarity of the vowel systems of all the known members of the Suto-Chuana group of Bantu languages" opens up the possibility of taking Southern Sesotho as representative of the others. On this basis the work of Doke and Mofokeng (1957) will be taken as representative of the rest.

Doke and Mofokeng (1957) provide a general outline in terms of articulation (Chapter 1). I will refer to this description, at the same time testing the validity of some specific descriptions portrayed as uncontested in all subsequent accounts of Southern Sotho, and likewise so in descriptions of members of the other Sotho languages. For example, in the case of Southern Sotho tone is said to distinguish grammatical and lexical meaning in all of the languages of the Sotho-Tswana group. While low tones can exist next to each other, high tones cannot (but see reference to Doke and Mofokeng's 1957 description further down below). In order to ensure maximum comparability I focus on one vowel only; primarily the half-close front vowel  $\bar{e}^1$ , situated between Vowel 1 and Vowel 2 on IPA's cardinal vowel chart. The focus word is the subjectival concord formative  $ke^2$ , in *ke motho* (I am a person). This will be done in contrast with the copulative *ke*, also in *ke motho* (He/she/it is a person). I shall also briefly go into possible tonal differences between subjectival concord formatives *ke* in the minimal pair *ke tēna* in two senses of 'I am dressing' and 'I am getting fed-up'. According to Doke and Mofokeng's (1957:46; 48; 302) description *ke* has a low tone in the present tense (positive), but is high in the negative. As maintained by these authors, such description applies to all speakers. I will put to the test this description as well as the assumption concerning the invariability as to individual speakers.

### 3. THE EMPIRICAL STUDY

The experiment comprises the elicitation and acoustic analysis of the production of the subjectival concord formative *ke*, as well as copulative *ke*. I focus on vowel quality and tone as the acoustic features.

#### 3.1. A note on Tone and Vowel Quality

While it is customary in phonology to refer to the phenomenon under discussion as *tone*, one should be sensitive to the relationship between *tone* and *pitch*. The latter denotes a perceptual property of a sound, perceived as for example *high* or *low*, and corresponds very closely to its fundamental frequency (F0). Pitch, therefore, is not an objective physical property of a sound, but a subjective psychophysical attribute of a sound. Tone is "made up" mainly of pitch, but vowel quality, intensity and even duration could contribute to a specific tonal quality too. In keeping with current practice, I will associate tone with pitch, and use F0 to quantify tonal values. Vowel quality will be expressed in terms of the first vowel formant, F1. Note here that F2, the indicator of horizontal position of vowels – e.g. front or back – is not relevant in the present study.

### 3.2. General Information

The deep-rural region of Phuthaditshaba, Eastern Free State, South Africa was chosen as research location for the first survey, and the nearby town Clarens for the second. Indications are that these are both very stable regions in terms of migrant patterns of the inhabitants.

### 3.3. Recordings

Due to the deep-rural situation in which some of the recordings had to be done, no technological support was available, such as PowerPoint screens on which the stimuli could be presented. Consequently I had to do with sheets of paper on which the reading materials were written. In some cases a battery-driven recorder was used, and in the rest a modern laptop. In the former case I recorded the speech samples using an Olympus WS-210S digital voice recorder. The recordings were sampled at standard 44100Hz, 16-bit, mono, in a Windows Media Audio bitrate format, 64 kbps, and then, via Adobe Audition converted to normal wave format (\*.wav), also with sampling frequency 44100 Hz, encoded in 16 bit linear PCM. In the second instance the same technical specifications were applicable, except that the recordings were done directly in wave format.

Similar recordings compared to studio recordings furnished acceptable results, so that I was confident in proceeding with the research as planned and described in this section<sup>3</sup>. Pitch (measured in terms of F0; tone), the single most important acoustic feature for the present research, is known to be quite consistent as to different recording equipment and environment. Formant frequency of the vowels, especially F1, the main indicator of vowel quality, also rendered very good consistency. I took all possible precautions to exclude external noises and preclude excessive reverberation. All other standard procedures were adhered to, such as keeping constant the distance between recorder and speaker's mouth, and controlling loudness of speech.

## 4. SPEECH PROCESSING

Oscillograms and spectrograms were produced in *Praat* (Boersma and Weenink, 2010). Speech signals were inspected auditorily and visually for determining vowel boundaries. Conventional segmentation criteria were followed (Grabe and Low 2002; Rietveld & Van Heuven 1997:108).

Vowels were annotated in *Praat*<sup>4</sup>, in such a way that *Vowelyse* (Van der Walt & Wissing 2004) could extract and calculate acoustic information relevant for analysis, which is, as already mentioned, first vowel formant (F1) and fundamental frequency (F0). Other information (e.g. duration and intensity) was also extracted.

Spurious formants and tone values are automatically detected and flagged by *Vowelyse* for later evaluation and possible removal after inspection and decision by the researcher. This is the case especially where *Praat* experiences problems in establishing a specific formant track, and then replaces the corresponding values with those of the next formant. In doing so, for example, one ends up with improbable F1 values like 2400 Hz (which is normal for F2) for *e* instead of round about 450 Hz.

#### 4.1. Participants of the Survey

Sixteen speakers of Southern Sotho were recruited for participation, eight of each gender, balanced in respect of age, thus rendering eight readers aged 55 years or older<sup>5</sup>, as well as eight of 15 – 17 years old. The latter group of participants were learners of a local high school. I will refer to the first groups as Adults, and to the second as Learners.

#### 4.2. Speech Stimuli

Two sentences each were constructed, (corresponding to 1 and 4 beneath), rendering the following:

3<sup>rd</sup> person copulative:

1. *Ke<sup>6</sup> motho* ("It is a person")
2. *Ke<sup>7</sup> nahana hore ke motho* ("I think it is a person")

1<sup>st</sup> person singular indicative present:

3. *Ke motho* ("I am a person")
4. *Ke nahana hore ke motho* ("I think I am a person")

In the case of 1 and 3 the English meaning were provided (resp. "It is a person", and "I am a person"). The English had to be read aloud as well. In case 2 it was indicated to the readers that *somebody was implied (the proper name Dineo was indicated in brackets, as was the pronoun nna in case 4)*.

In constructing the stimuli, I thus took care of controlling varying possible phonetic influences. This includes position in the sentence, and a possible consonantal influence (by voiced plosives *b* or *d*) on tonal behaviour. Grammatical differences between the two kinds of *ke*'s were not deemed important in a study of phonetic features as the present one.

Each speaker read each sentence three times, rendering 192 sentences. As I was interested in the *ke* of *ke motho* in each of the four sentences, ideally I had access to 192 *e* vowels. Though of secondary interest, I included the first<sup>8</sup> vowel of *motho* in

the speech processing (N = 192). In actual fact, these numbers were lower, due to a variety of reasons, among others unclear recordings.

## 5. RESULTS

As was stated in the Introduction, Doke and Mofokeng (1957) and others describe *e* of first person subjectival concord formative *ke* (*ke1*) as being of low tone, in contrast with *e* of the third person copulative *ke* (*ke3*), which is taken to be high in tone. According to this (and other) work no difference in terms of vowel quality was to be expected.

In the following account KE (an abbreviation referring to a combination of *ke1* and *ke3*) is the main categorical predictor (factor), with AGE and GENDER the other factors (independent variables). AGE refers to either Adult readers or learners, and GENDER to Male or Female. TONE and VOWEL QUALITY respectively<sup>9</sup> are the dependent variables, both measured in terms of formant frequency. In this regard I made use of analysis of variance (ANOVA), which enables tests for significant differences between means. In some cases a one-way ANOVA was sufficient. A one-way ANOVA is used to analyse designs with a single categorical independent factor (variable). In more complex designs main effects ANOVAs were used. The latter are used to analyse the first-order (non-interactive) effects of multiple categorical dependent variables, in the present case AGE and GENDER.

Great similarity in terms of the two independent variables of *ke* in the four sentences mentioned above justified the pooling together of two instances (Sentence 1 and 3) of the 1st person subjectival concord formative *ke* as one class, henceforth *ke1*, and likewise the two instances (Sentence 1 and 3) of the 3rd person copulative morpheme *ke* as a second class, *ke3*. In the following tables the relevant information with regard to the independent variables KE, GENDER and AGE will be presented.

In Table 1 – Table 4 the basic measurements of tone and vowel quality, in terms of F0 and F1 respectively (both given in Hertz), are presented.

**Table 1:** *ke1* and *ke3*, all cases pooled

Morpheme	F0 mean	F0 N	F0 s.d.	F1 mean	F1 N	F1 s.d.
Ke1	170	85	47.6	453	87	64.8
Ke3	181	94	55.7	456	92	58.7



**Table 2:** *ke1* and *ke3*, per Gender

<b>GENDER Morpheme</b>		<b>F0 mean</b>	<b>F0 N</b>	<b>F0 s.d.</b>	<b>F1 mean</b>	<b>F1 N</b>	<b>F1 s.d.</b>
Male	Ke1	127	41	25.5	426	46	39.3
Male	Ke3	146	49	52.0	438	50	54.0
Female	Ke1	210	44	20.8	484	41	73.7
Female	Ke3	219	45	27.4	478	42	56.9

**Table 3:** Table 2: *ke1* and *ke3*, per Age.

<b>AGE</b>	<b>Morpheme</b>	<b>F0 mean</b>	<b>F0 N</b>	<b>F0 s.d.</b>	<b>F1 mean</b>	<b>F1 N</b>	<b>F1 s.d.</b>
Learners	Ke1	172	45	51.9	467	46	75.6
Learners	Ke3	177	46	57.1	458	46	73.3
Adults	Ke1	167	40	42.7	438	41	45.9
Adults	Ke3	185	48	54.6	454	46	39.8

**Table 4:** Table 2: *ke1* and *ke3*, per Gender and Age.

<b>AGE + GENDER</b>	<b>Morpheme</b>	<b>F0 mean</b>	<b>F0 N</b>	<b>F0 s.d.</b>	<b>F1 mean</b>	<b>F1 N</b>	<b>F1 s.d.</b>
Adult female	Ke1	197	22	15.8	445	19	57.4
Adult female	Ke3	204	23	27.8	444	20	48.1
Learner female	Ke1	223	22	16.4	518	22	70.3
Learner female	Ke3	235	22	16.6	510	22	45.3
Adult male	Ke1	131	18	36.9	431	22	33.2
Adult male	Ke3	166	25	66.4	463	26	30.8
Learner male	Ke1	124	23	10.1	421	24	44.2
Learner male	Ke3	125	24	11.8	411	24	61.1

In Table 5 I present results of ANOVA analyses with regard to the specific values as mentioned in Table 1 – 4. One-way ANOVAs were utilised in the case of the first three tables, while main effects ANOVAs were done for complex designs (Table 4). In all cases Such ANOVAs were applied to the product of a process of statistical normalisation (also called standardisation).<sup>10</sup> The following exposition and

discussion of the results in Tables 5 should be read in close proximity with the general information of the first four tables.

**Table 5:** Comparisons of *ke1* with *ke3* of Gender and Age groups as well as combined groups

<b>GENDER</b>	<b>p-value F0</b>	<b>p-value F1</b>
Male	F(1, 88)=10.4, p=.002	F(1, 94)=.002, p=.97
Female	F(1, 87)=5.8, p=.02	F(1, 81)=.01, p=.92
<b>AGE</b>		
Adults	F(1, 86)=14.5, p=.0003	F(1, 85)=4.7, p=.03
Learners	F(1, 89)=3.5, p=.06	F(1, 90)=1.4, p=.23
<b>GENDER + AGE</b>		
Adult males	F(1, 41)=25.9, p=.00001	F(1, 46)=14.3, p=.0005
Adult females	F(1, 43)=1.7, p=.2	F(1, 37)=.23, p=.6
Learner males	F(1, 45)=.2, p=.7	F(1, 46)=1.4, p=.2
Learner females	F(1, 42)=5.3, p=.03	F(1, 42)=.2, p=.7

### 5.1. Results Concerning Tone (F0):

While it is quite clear that according to the results of Table 5 tone indeed does play a role in most cases, especially in bringing about a significant difference between *ke1* and *ke3*, the strong position of adult males in particular (p=.00001) is striking. Noteworthy is the insignificant difference in the production adult females' morphemes *ke1* and *ke3* (p=.2), and also the difference between Adults and Learners (Learner males contribute largely to this). These results thus only partially support the present description of the role of tone in the case of *ke1* and *ke3*, viz. that *ke1* is low in tone, *ke3* high. What needs to be observed is the direction of F0 in these two morphemes: in all case but one (Learner males) *ke3* is produced higher in tone, in accordance with existing descriptions. Whether such higher tone values would suffice in the interpretation of such cases remains to be seen. An assisting role of vowel quality here does not seem unlikely. Research in the form of perception tests is necessary.

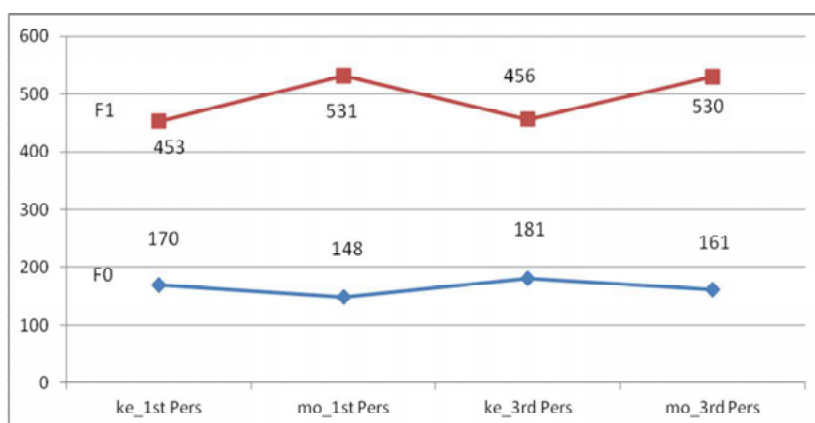
### 5.2. Results Concerning Vowel Quality (F1):

At first sight the results as to the role of vowel quality in producing *ke1* and *ke3* seem to support the current descriptions in that the /e/ vowel is identically pronounced with all cases pooled. Closer inspection reveals the exception in the case of Adults (p=.03), but more strikingly Adult males' production (p=.0005), indicative of a distinct control over this aspect. However, what needs to be pointed

out is the fact that this group of speakers' specific F1 readings lie in the opposite direction to what would have been expected ( $ke1 = 431$  Hz;  $ke3 = 463$  Hz). Taking into account the general acoustic phonetic viewpoint that there is a direct relation between vowel quality and tone (higher vowels tend to be produced with higher tone, see Kent & Read 1992:95) in this case the expectation of a reverse situation is more probable and to be expected, which would rather mean  $ke1 = 463$  Hz;  $ke3 = 431$  Hz. This could be interpreted as a lack of control by these readers, or a counterexample of such a relation between vowel quality and tone as posed by general phonetic theory. More research with regard to this aspect is needed.

A note on *mo-* perhaps is necessary here. As seen above, *mo-* should always be low in tone, and with invariable vowel quality. Concerning the latter, Figure 1 demonstrates that this is indeed the case. On the other hand, as to its tone, *mo-* in the 3<sup>rd</sup> person (161 Hz) is notably higher than *mo-* in the 1<sup>st</sup> person (148 Hz). This may be ascribed to influence of preceding high F0 of 1<sup>st</sup> person *ke*, or it might be possible that such a raised tone on this *mo-* could serve as strengthening factor of *ke*. This possibility can be investigated in a perception study in which both *motho* structures (1<sup>st</sup> and 3<sup>rd</sup> person) without *ke* are presented to listeners for them to discriminate between the two, and, importantly, to identify which is which. Such a study is at the moment ongoing<sup>11</sup>.

**Figure 1:** F1 and F0 measurements of the vowels in *ke motho* in 1<sup>st</sup> and 3<sup>rd</sup> person. Y-axis denotes Hertz.



In summary, these findings only partly confirm the view of Doke and Mofokeng (1957) concerning the uniformity of tonal behaviour as well as the invariability of vowel quality in the speech production of Sotho speakers. Overall, the results of this survey suggest a cautious acceptance of their view concerning the consistent and invariant production of instances such as these.

### 5.3. *Ke* as Subjectival Concord Formative

Ten learners aged on average 14 years, five of each gender, volunteered as readers. Comparison of the measurements the production of the two vowels *ke* and *ten-* in *ke tena*1 (“to get dressed” versus *ke tena*2 (“to get fed-up), allowed me to merge these two constructions’ vowels to a single one for the positive, viz. *ke tena*, and one for the negative, viz. *ke tene*. According to Doke and Mofokeng (1957:46) *ke* is high-toned in the negative mode; low in the positive, with invariable vowel quality, just like in the case of the vowel of *ke*1 and *ke*2 of the previously reported experiment, viz. “ē” in their transcription, referring to the mid-high vowel – situated between Cardinal Vowel 1 and 2.

In summary I found only slight support for Doke and Mofokeng’s (1957:46) description of tonal behaviour of *ke* in these two structures. F0 for the positive is 203 Hz; that of the negative 211 Hz ( $p=0.05$ ), indicating only a slight raise of tone in the negative. On first sight both could be characterised as high in tone, rather than the expected low in the case of the positive *ke*, high for negative *ke*. On the other hand, in the case of F1 the difference is statistically highly significant (positive: 424 Hz; negative 351 Hz) ( $p=0.0000$ ).<sup>12</sup> Importantly the *ke* of the indicative negative form’s reading of 351 Hz “correctly” indicates a higher vowel than the positive *ke*, in comparison with the “wrong”<sup>13</sup> direction found in the case of Adult males in the previous section. Here, however, it should be kept in mind that only young speakers were involved.

**Table 6:** Subjectival concord formative *ke* in both structures. All groups (male + female Learners) pooled.

STRUCTURE	F0 mean	F0 N	F0 s.d.	F1 mean	F1 N	F1 s.d.
IND_POS <i>ke</i> (a <i>tena</i> )	203	31	52	424	31	49.6
IND_NEG (ha) <i>ke</i> ( <i>tene</i> )	211	32	55	351	31	66.8

**Table 7:** Results of male and female Learners concerning the subjectival concord formative *ke* in both structures

GENDER	STRUCTURE	F0 mean	F0 N	F0 s.d.	F1 mean	F1 N	F1 s.d.
Female	IND_POS_ <b>ke</b> (a <i>tena</i> )	232	21	35.1	441	21	46.9
Female	IND_NEG_(ha) <b>ke</b> ( <i>tene</i> )	240	22	36.9	376	21	67.8
Male	IND_POS_ <b>ke</b> (a <i>tena</i> )	142	10	12.6	389	10	35.3
Male	IND_NEG_(ha) <b>ke</b> ( <i>tene</i> )	145	10	16.6	299	10	15.6

In Table 8 I give the statistics in terms of p-values for Male and Female speakers combined as well as separately.

**Table 8:** Comparisons of *ke* in positive versus indicative negative structures as produced by Male and Female speakers.

<b>ALL GROUPS</b>	<b>p-value F0</b>	<b>p-value F1</b>
Male + Female	F(1,61) = 6.5, p = 0.01	F(1,60) = 34.9, p = 0.00000
<b>GENDER</b>		
Male	F(1,18) = 1.4, p = 0.3	F(1,18) = 32.8, p = 0.00002
Female	F(1,41) = 5.1, p = 0.03	F(1,40) = 159, p = 0.0003

Broadly taken the results of Table 8 (again to be read in close connection with those of Table 6 and 7) are again, as in the case of the first study, in line with the general expectation: *ke* of *ke tena* 1(indicative positive) is significantly lower in tone than *ke tene* 2 (F0: p=.01). Specifically note, however, the insignificant difference made by the male Learners (F0: p=.3). Furthermore, contrary to existing descriptions it is clear that the quality in terms of F1 of /e/ vowel of both *ke*'s is distinctly dissimilar (F1: p=.00000). This largely also goes for male and female speakers taken separately (F1:p<.0004).

Probably the notable difference in vowel quality could, at least partly, be attributable to coarticulatory influence of the following stem vowel of negative *tenē* (F1 = 354 Hz), in its turn being raised under influence of the word final –e of negative *tenē* (F1 = 328 Hz). Whatever the case may be, this phenomenon was notably not taken into account by Doke and Mofokeng (1957) in their description as cited above (cf. Wissing 2010b for an analysis of vowel raising).

The slightly higher tone of the negative *ke* could to a certain extent be explained by a factor other than its function of expression of the negative, as held in existing descriptions. It may be possible that the slightly higher tone of *ke* here is caused by the well-known phenomenon of inherent fundamental frequency (F0) (Kent and Read, 1992:95). These authors cite a difference of 5% - 12% between mid front /ɛ/ to high front /i/ (vowels nearest to those of *ke tena* and *ke tene*). Taking this into account, it seems that most of the tonal differences found here could be expected in addition to, or instead of a primary grammatical purpose of expression of tone.

The overall picture is clear: neither tone nor vowel quality can be seen in the uniform way existing textbooks do.

## 6. CONCLUSION AND PROSPECT OF FUTURE RESEARCH

Generally the findings of this investigation suggest the lack of complete control of tone by different groups of Sesotho speakers. This means that the currently held belief that tone is a definite and robust indicator of grammatical differences between morphemes such as *ke* cannot be fully supported. Only in the case of adult males this seems to be true. What is fascinating about this, is the possibility of transfer of the command over tone from mother to child. This explanation tallies rather well with known sociolinguistic belief (cf. e.g. Labov 2002). There seems to be a sociolinguistic explanation for the findings with respect to the differences between especially the male adults and the rest. Should one accept the validity of the term “**mother-tongue**” for “first language”, it may be possible that the females (**mothers**), who have lost the tone contrast have transferred that to the next generation, here the young speakers, especially the boys. Females, here the mothers in the older generation, according to Labov (2002), generally are at the forefront of many types of phonetic change such as these<sup>14</sup>.

It also is a possibility that grammarians such as Doke and Mofokeng (1957), in classifying *ke* of the indicative negative *ke tene* as of high tone, were led to do so erroneously on the grounds of the perceptual impression of high vowel quality in combination with an only slightly higher tone, such higher tone being reinforced by a significantly different vowel quality. This explanation will have to be investigated more closely by keeping in mind verbs with vowel stems other than –e-.

Perhaps Roux (1981) was justified in proposing his Perceptual Confusion Hypothesis, based on precisely such confusion on the side of the listener in the sense of what Tucker (1929:23) referred to with his example of *ke tau* (cf. Introduction), namely that high tone may produce the effect of higher vowels. Of course it may be possible that this applies the other way around too: high vowels may produce the impression of high(er) tone. Finally, a combination of slightly higher tone with high vowels may induce the effect of a considerably higher tone.

This is a topic in the sound system of Sesotho that needs to be researched in detail. In this regard perception tests are essential. Utilising production samples of some of the adult speakers as well as those of younger persons, especially male speakers, could be of special interest. In terms of the topic of language change, findings in this regard may be significant in that it could suggest a loss of tone contrast in at least some structures, for instance subjectival concord formatives and copulatives.

## ENDNOTES

1. Doke and Mofokeng's (1957) transcription – see Note 3.
2. Except in the case of explicit quotes, in future *kē* and *mōthō* will be written simply as *ke* and *motho*. Doke and Mofokeng (1957) use the transcription with *ē* and *ō*.
3. During a lecture by myself on this topic (the University of Michigan, Department of Linguistics' Research Group: Phonetics and Phonology), consensus as to the acceptability was reached, adding confidence in such methodology.
4. Praat is a comprehensive speech analysis programme, and Vowelyse is an add-on script specialised for the analysis of vowels.
5. The ideal was to record much older persons, but that turned out not to be practically possible.
6. Here *ke* is the subjectival concord formative.
7. Here *ke* is the copulative morpheme.
8. Excessive dampening was present in the second vowel of "o" in *mōthō*, which is quite normal in the case of low tone vowel in phrase-final position.
9. Measured in terms of F0 and F1 respectively. For a short explanation of these terms, see lower down.
10. The process of normalisation refers to the transformation of data by subtracting some reference value (typically a sample mean) from each value and dividing it by the standard deviation (typically a sample s.d.). This important transformation will bring all values (regardless of their distributions and original units of measurement) to compatible units from a distribution with a mean of 0 and a standard deviation of 1. This transformation makes the distributions of values easy to compare across variables and/or subsets. (Extraction from Statistica's Glossary. Item: Standardization).
11. Collaborators are Justus Roux and Andries Coetzee.
12. Here too, as previously, only ANOVAs were done on normalised values.
13. I provisionally use the words in quotation marks in the sense of Doke and Mofokeng's (1957) description.
14. I would like to thank Bertus van Rooy for drawing my attention to this facet.

## REFERENCES

- Barnard, E. & D.P. Wissing. 2008. Vowel variation in Southern Sotho: an acoustic investigation. *Southern African Linguistics and Applied Language Studies* 26(2):255-265.
- Boersma, P. & D. Weenink, *Praat, a system for doing phonetics by computer*. Version 5.1.43. <http://www.fon.hum.uva.nl/praat>. Accessed: 10-10-2010.
- Cole, D. T. 1955. *An Introduction to Tswana Grammar*. Cape Town: Longman.
- Doke, C.M. & S.M. Mofokeng. 1957. *Textbook of Southern Sotho Grammar*. London/ Cape Town/ New York: Longman.
- Grabe, E. & E. Ling Low. 2002. Durational variability in speech and the rhythm class hypothesis. *Laboratory Phonology VII*, edited by Carlos Gussenhoven and Natasha Warner. Berlin: Mouton de Gruyter. 515-646.
- Guma, S.M. 1971. *An outline structure of Southern Sotho*. Pietermaritzburg: Shuter and Shooter.
- Khabanyane, K.E. 1991. *The five phonemic vowel heights of Southern Sotho: an acoustic and phonological analysis*. Working papers of the Cornell Phonetics Laboratory, No. 5, Sept. 1991.
- Krüger, C.J.H. & J.W. Snyman. s.a. *The sound system of Setswana*. Goodwood: Via Africa.
- Poulos, G. & L.J. Louwrens. 1994. *A linguistic analysis of Northern Sotho*. Goodwood: Via Africa.
- Labov, W. 2002. *Principles of linguistic change. Volume 1, Internal factors*. Oxford and Cambridge: Blackwell.
- Letele, G.L. 1955. *The role of tone in the Southern Sotho language*. Johannesburg: Printed.
- Rietveld, A. & V. van Heuven. 1997. *Algemene fonetiek*. Bussum: Coutinho.
- Roux, J.C. 1979. *Labialization in Sesotho: The role of phonetic data in phonological analyses*. Unpublished Doctoral Dissertation. Stellenbosch: University of Stellenbosch.
- Roux, J.C. 1983. *Vokaalkwaliteit, toon en konsonantale invloede in Sotho*. Unpublished research report. Stellenbosch: University of Stellenbosch.
- Selebeleng, K.Z. 1997. *Phonetic and phonological aspects of vowel raising in SeSotho*. MA thesis. Stellenbosch: University of Stellenbosch.



- Tucker, A.N. 1929. *The comparative phonetics of the Suto-Chuana group of Bantu languages*. London: Longman, Green and Co.
- Van der Walt, A.J. & D.P. Wissing. 2004. *Vowelyse: tools for acoustic analysis and optimising phonetic actions*. [Computer Programme].  
[http://www.puk.ac.za/fakulteite/lettere/HLT\\_resources](http://www.puk.ac.za/fakulteite/lettere/HLT_resources).
- Wissing, D.P. 2005. Aspiration of English voiceless stop consonants in Southern Sotho: a case study. *South African Journal of African Languages* 25(3):189-205.
- Wissing, D.P. 2010a. Aspects of the phonetics and phonology of Southern Sotho /a/. *South African Journal of African Languages* 30(2):234-241.
- Wissing, D.P. 2010b. Aspects of vowel raising in Southern Sotho and Setswana: An acoustic approach. *South African Journal of African Languages* 30(2):242-249.

## CHAPTER 5

### SYLLABIFICATION OF CONSONANTS IN SESOTHO AND SETSWANA

Mildred Nkolola-Wakumelo<sup>1</sup>, Liketso Rantso<sup>2</sup>, Keneilwe Matlhaku<sup>3</sup>

<sup>1</sup>*Department of African Languages and Literature, University of Zambia  
Department of African Languages, University of Witwatersrand, South Africa  
mwakumelo@yahoo.com*

<sup>2</sup>*Department of African Languages and Literature National University of Lesotho  
flrantso@yahoo.co.uk*

<sup>3</sup>*Department of African Languages and Literature, University of Botswana  
maswabik@mopipi.ub.bw*

#### 1. INTRODUCTION

This paper discusses the occurrence of derivative syllabic consonants in Sesotho and Setswana. Sesotho and Setswana belong to the group of languages generally referred to as the Sotho-Tswana languages. Guthrie (1948) groups these languages under Zone S30. In our paper we do not discuss the syllabic consonants in Sesotho and Setswana for comparative purposes.

Instead we discuss this phenomenon with reference to the two languages for the purpose of showing that the process is characteristic of these two languages both belonging to the Sotho-Tswana group. The theoretical perspectives applied in our discussion are drawn from Distinctive Feature Phonology, proposed by Jakobson *et al.* (1952) and further developed by Chomsky and Halle (1968).

Using this approach we present and explain the phonological processes involved in the syllabification of consonants using features. One of the major principles of Distinctive Feature Phonology is that phonemes are regarded as bundles of features represented with a binary system where [+] means a feature is present and [-] means that a feature is absent. According to Chomsky and Halle (1968) distinctive features are designed to describe the phonetic content of segments derived by phonological rules, as well as underlying segments (refer also to Hyman 1975).

In addition Chomsky and Halle (1968) state that these features capture natural classes, which are classes of sounds that share features and would tend to behave the same way in the same environment. It is also stipulated in this theory that some phonological rules apply to classes of phonetically related sounds (that is, sounds that form a natural class) (refer also to O'Grady *et al.* 1997).

## 2. DATA COLLECTION METHODS

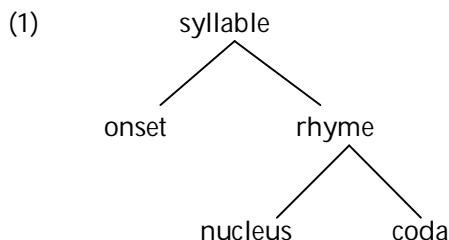
The data that is used in our discussion of the derivative syllabic consonants was obtained from various sources. Some of the data was collected from native speakers of the languages under discussion while some of it was obtained from various written sources of the two languages. The native speakers of the languages that were consulted during the compilation of the data were based in Lesotho (for Sesotho) and Botswana (for Setswana). Their age range was from 40 years upwards. The choice of this age range was based on the premise that the older generation will in most cases speak the purer forms of languages. It should also be pointed out that two of the writers of this chapter are native speakers of Setswana and Sesotho and were therefore able to provide very useful insights on the validity of the data collected and its analysis. The importance of native speaker intuitions is well-known among linguists. On this Atkinson *et al.* (1982:38) point out that if

[t]he linguist is a native speaker of the language he is investigating he will be able to distinguish between well-formed and ill-formed strings of words... and is entitled to invent sentences and non-sentences to formulate and test his hypotheses.

Atkinson *et al.* (1982) add that such abilities of a native speaker of a language are what are known as linguistic intuitions which form an essential part of the database of a Chomskyan approach to linguistics which will contain not only utterances but judgments about such utterances. Apart from data collected from native speakers of the two languages the writers also consulted written sources in the two languages. These included grammars, dictionaries, readers and school textbooks. The written sources were consulted to avoid falling into an unconscious bias consisting in looking for data which suits one's objectives. Data collection and analysis were carried out in phases as follows. Firstly, the writers collected a substantial amount of data on instances of the occurrence of syllabic consonants from written sources of the two languages. Secondly, this data was verified in consultation with native speakers of the languages. Thirdly, using two of the writers' knowledge of the languages as native speakers, they checked the data collected, eliminating those that were dubious. The fourth phase consisted of sorting out the data and identifying situations of the occurrence of derivative syllabic consonants.

## 3. IDENTIFYING THE SYLLABLES IN SESOTHO AND SETSWANA

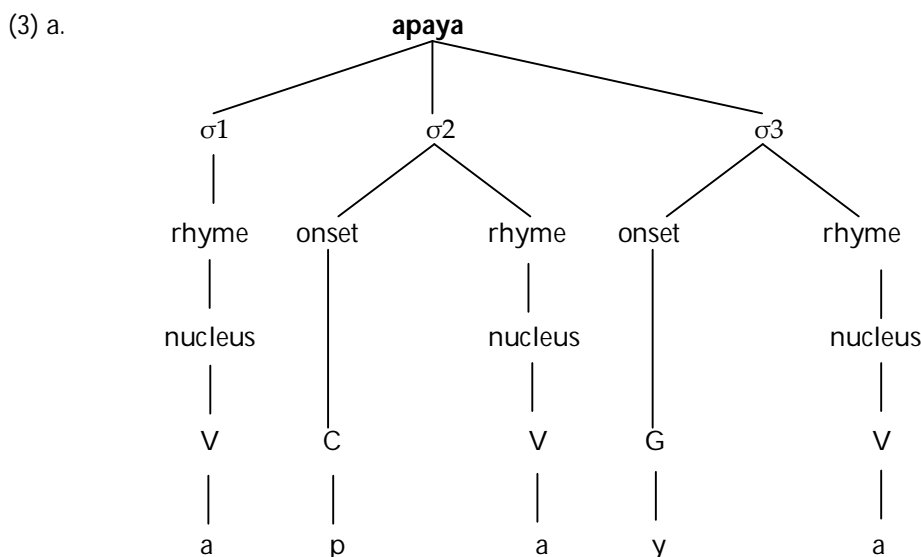
The general structure of a syllable as proposed by Cairns and Feinstein (1982) is as illustrated in (1) below:

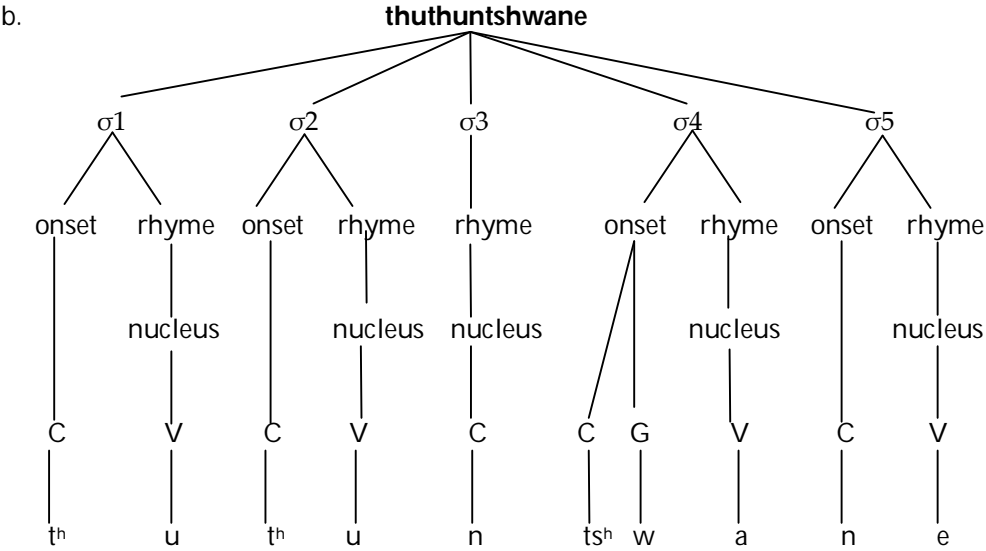


The tree-diagram in (1) is derived from the following rule:

- (2) (a) Syllable = Onset + Rhyme  
 (b) Rhyme = Nucleus + Coda

Both Sesotho and Setswana like other Bantu languages have open syllables. In Sesotho and Setswana the syllable can be made up of different structures. As illustrated using the words **apaya** (a.pa.ya) 'cook' and **thuthuntshwane** (thu.thu.n.tshwa.ne) 'a wild mushroom/fungus' below, the Sesotho and Setswana syllable structure can be made up of a consonant + glide + vowel (CGV) as in syllable 4 ( $\sigma_4$ ) in 3(b); consonant + vowel (CV) as in syllable 2 ( $\sigma_2$ ) in 3(a), syllables 1, 2 and 5 ( $\sigma_1$ ,  $\sigma_2$  and  $\sigma_5$ ) in 3(b); consonant (C) as in syllable 3 ( $\sigma_3$ ) in 3(b); glide + vowel (GV) as in syllable 3 ( $\sigma_3$ ) in 3(a); or vowel (V) as in syllable 1 ( $\sigma_1$ ) in 3(a). On the other hand the onset can be made up of the structures consonant and glide (CG) (cf. syllable 4 ( $\sigma_4$ ) in 3(b)); consonant (C) (cf. syllable 2 ( $\sigma_2$ ) in 3(a) and syllables 1, 2 and 5 ( $\sigma_1$ ,  $\sigma_2$  and  $\sigma_5$ ) in 3(b)); or glide (G) (cf. syllable 3 ( $\sigma_3$ ) in 3(a)).





**4. OCCURRENCE OF SYLLABIC CONSONANTS IN SESOTHO AND SETSWANA**

A consonant which constitutes a syllable is said to be syllabic. It is one which can form a syllable on its own or constitutes the nucleus of a syllable. Sesotho and Setswana are some of the Bantu languages which have syllabic consonants which can be nasals and liquids. Sesotho has five syllabic consonants; these include the nasal sounds [m], [n], [ɲ] and [ŋ] and the liquid [l] (Guma 1971). On the other hand Setswana has six syllabic consonants. These are the nasals [m], [n], [ɲ], [ŋ], the liquid [l] and the trill [r] (Cole 1955). We illustrate these below.

(4)

**(a) Syllabic [m]**

[m.p <sup>h</sup> ɔ]	mp <sup>h</sup> o	'gift'	← fa 'give'	(Sesotho and Setswana)
[m.pa]	mpa	'stomach'		(Sesotho and Setswana)
[m.mɛ]	'me	'mother'		(Sesotho and Setswana)
[m.pi.tsa]	mpitsa	'call me'	← bitsa 'call'	(Sesotho and Setswana)
[m.p <sup>h</sup> a]	mp <sup>h</sup> a	'give me'	← fa 'give'	(Sesotho and Setswana)
[m.pɔ.na]	mpona	'see me'	← bona 'see'	(Sesotho and Setswana)

**(b) Syllabic [n]**

[n.ts <sup>h</sup> i]	ntši	‘eyelash’	(Sesotho and Setswana)
[n.ta.tɛ]	ntate	‘father’	(Sesotho and Setswana)
[n.tlo]	ntlo	‘house’	(Sesotho and Setswana)
[n.thɔ]	ntho	‘wound’	(Setswana)
[n.na]	nna	‘me/ I’	(Sesotho and Setswana)
[n.tʃa]	ntša	‘dog’	(Setswana)

**(c) Syllabic [ɲ]**

[ɲ.tʃa]	ntja	‘dog’	(Sesotho)
[ɲ.tʃ <sup>h</sup> a]	ncha/ntšha	‘new’	(Sesotho and Setswana)
[ɲ.ɲe]	nnye	‘small’	(Setswana)

**(d) Syllabic [ŋ]**

[ŋ.ku]	nku	‘sheep’	(Sesotho and Setswana)
[ŋ.q <sup>h</sup> a]	nkga	‘smell’	(Setswana)

**(e) Syllabic [l]**

[mo.l.lɔ]	mollo	‘fire’	(Sesotho and Setswana)
[so.l.la]	solla	‘wander’	(Sesotho)
[bo.fu.l.la]	bofolla	‘untie’	(Sesotho and Setswana)
[si.l.lɔ]	sello	‘a cry’	(Sesotho and Setswana)

**(g) Syllabic [r]**

[r.ra]	rra	‘Mr.’	(Setswana)
--------	-----	-------	------------

In both Sesotho and Setswana consonants can be syllabic in non-derivative and derivative contexts. Hence we can make a distinction between non-derivative and derivative syllabic consonants. Derivative syllabic consonants are consonants that were originally not syllabic that become syllabic. In the languages under discussion it has also been noticed that in derivative contexts the occurrence of the syllabic consonants can be due to some phonological process which may be induced by the

affixation of some inflectional or grammatical morpheme to a word. These consonants can be syllabified in word initial, medial and final positions.

According to Martin (2003) the process of making consonants syllabic is called syllabification. On the other hand Kunene (1961) states that syllabification is the process of the acquisition of syllabicity by a consonant which was initially non-syllabic. Following Kunene (1961) and Martin (2003) we have also referred to the realisation of derivative syllabic consonants as the process of syllabification. According to Kunene (1961) in syllabification the accompanying change is the elision of a vowel immediately following the affected consonant. She also notes that syllabification occurs when "a vowel separating two l's, m's or n's is elided". Other scholars such as Guma (1971), Doke and Mofokeng (1985) also mention phonological processes such as elision as being the processes that are involved in the syllabification of consonants. While all the above scholars mention elision as the main process to which they attribute the syllabification of consonants, our analysis has shown that there are other phonological processes that bring about the syllabification of consonants.

Our analysis of data from the two languages has shown that the phonological processes that are involved are deletion, place of articulation assimilation and manner of articulation assimilation. We have also referred to the latter two processes as homorganic nasalisation and nasalisation, respectively. However, it should be pointed out that the derivative contexts of the syllabic consonants we discuss below may not be the only ones in the languages concerned. We have chosen to discuss the contexts below mainly because they exhibit a widespread phenomenon of the syllabification of consonants in the languages. For instance, it is possible for [m] to occur as a derivative syllabic consonant with the prefixation of the nominal derivation prefix **mo-** to some verbs in the formation of deverbal nouns. This can be seen below:

- (5) bala → mo-bal-i → mbali → mmali [m.ma.di] (Sesotho)  
 'read' 'reader'  
 bala → mo-bal-I → mbadi → mmadi [m.ma.di] (Setswana)  
 'read' 'reader'  
 busa→mo-bus-i→ mobus-i → mbus-i→ mmusi [m.mu.si](Sesotho and Setswana)  
 'lead' 'leader'

In this contribution we have not discussed instances such as these because this phenomenon of the syllabification of [m] in this position is not widespread for all

agentive deverbals. It only affects a few deverbals in the languages. It occurs only when the initial consonant of the verb to which the prefix **mo-** is attached is the voiced bilabial plosive [b].

## 5. SYLLABIFICATION OF CONSONANTS

### 5.1 Syllabification of Consonants in Word Initial Position

In the derivative contexts the nasals [m], [n], [ɲ] and [ŋ] can be syllabic and may occur word initially as a result of the affixation of the first person singular object marker (OM) /n/ to some verbs in Sesotho and Setswana. The affixation of the object marker triggers place of articulation assimilation which causes the object marker [n] to change to [m], [ɲ] or [ŋ]. In this case these syllabic nasals are the realisations of the object marker prefix /n/ due to some phonological processes. This can be seen in the words below:

(6)

- |     |                            |   |                                   |
|-----|----------------------------|---|-----------------------------------|
| (a) | n + [p <sup>h</sup> ɛhɛla] | →*[np <sup>h</sup> ɛhɛla]→[m.phɛ.hɛ.la] | (mphehela) (Sesotho and Setswana) |
|     | OM 'cook'                  |   | 'cook for me'                     |
| (b) | n + [mɛma]                 | →*[nmɛma] →[m.mɛ.ma]                    | ('mema) (Sesotho)                 |
|     | OM 'invite'                |   | 'invite me'                       |
| (c) | n + [bona]                 | →*[nbona] →[m.po.na]                    | (mpona) (Sesotho and Setswana)    |
|     | OM 'see'                   |   | 'see me'                          |
| (d) | n + [tɪna]                 | → [n.tɪ.na]                             | (ntena) (Sesotho and Setswana)    |
|     | OM 'annoy'                 |   | 'annoy me'                        |
| (e) | n + [t <sup>h</sup> usa]   | → [n.t <sup>h</sup> u.sa]               | (nthusa) (Sesotho and Setswana)   |
|     | OM 'help'                  |   | 'help me'                         |
| (f) | n + [dia]                  | → *[ndia] →[n.ti.a]                     | (ntia) (Sesotho and Setswana)     |
|     | OM 'delay'                 |   | 'delay me'                        |
| (g) | n + [sia]                  | → *[nsia] → [n.ts <sup>h</sup> i.a]     | (ntshia) (Sesotho and Setswana)   |
|     | OM 'run'                   |   | 'leave me behind/ run from me'    |
| (h) | n + [tsɪba]                | → [n.tsɪ.ba]                            | (ntseba) (Sesotho)                |
|     | OM 'know'                  |   | 'know me'                         |
| (i) | n + [ts <sup>h</sup> ola]  | → [n.ts <sup>h</sup> o.la]              | (ntshola) (Sesotho and Setswana)  |
|     | OM 'give birth'            |   | 'give birth to me'                |



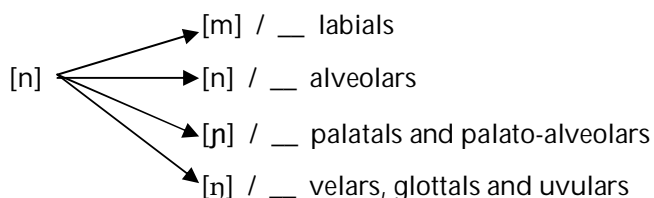
- (j) n + [tlola] → [n.tlɔ.la] (ntlola) (Sesotho and Setswana)  
OM 'jump over' 'jump over me'
- (k) n + [tlhapa] → [n.tlh̥a.pi.sa] (ntlhapisa) (Sesotho and Setswana)  
OM 'bathe' 'bathe me'
- (l) n + [luma] → \*[nluma] → [n.tɔ.ma] (ntoma) (Sesotho and Setswana)  
OM 'bite' 'bite me'
- (m) n + [rɔχa] → \*[nrɔχa] → [n.tʰɔ.χa] (nthoga) (Sesotho and Setswana)  
OM 'insult' 'insult me'
- (n) n + [rata] → \*[nrata] → [n.tʰa.ta] (nthata) (Sesotho and Setswana)  
OM 'love/ like' 'love/like me'
- (o) n + [lapa] → \*[nlapa] → [n.tʰa.pa] (nchapa/ntšhapa) (Sesotho and Setswana)  
OM 'beat' 'beat me'
- (p) n + [ɲala] → \*[ɲɲala] → [ɲ.ɲa.la] ('nyala) (Sesotho and Setswana)  
OM 'marry' 'marry me'
- (q) n + [ɲɛɲafatsa] → \*[ɲɛɲafatsa] → [ɲ.ɲɛ.ɲa.fa.tsa] (nnyenyafatsa) (Sesotho and Setswana)  
OM 'belittle' 'belittle me'
- (r) n + [dʒɪsa] → \*[ndʒɪsa] → [ɲ.dʒɪ.sa] (njesa) (Sesotho and Setswana)  
OM 'feed' 'feed me'
- (s) n + [kama] → \*[nkama] → [ɲ.ka.ma] (nkama) (Sesotho and Setswana)  
OM 'comb' 'comb me'
- (t) n + [qʰama] → [nqʰama] → [ɲ.qʰa.ma] (nkgama) (Sesotho and Setswana)  
OM 'strangle' 'strangle me'
- (u) n + [χɔχa] → \*[nχɔχa] → [ɲ.qʰɔ.χa] (nkgoga) (Sesotho and Setswana)  
OM 'pull' 'pull me'
- (v) n + [χana] → \*[nχana] → [ɲ.qʰa.na] (nkgana) (Sesotho and Setswana)  
OM + 'refuse' 'refuse me'
- (w) n + [hira] → \*[nhira] → [ɲ.kʰi.ra] (nkhira) (Sesotho and Setswana)  
OM 'hire' 'hire me'

The syllabification of the nasals [m], [n], [ɲ] and [ŋ] with the object marker /n/ involves place of articulation assimilation, whereby /n/ assimilates the place of articulation of the first consonants of the verbs to which it is attached. This, according to distinctive feature phonology happens because, when sound segments are in the neighbourhood of others, they tend to assimilate features of segments around them. This assimilation is regressive or anticipatory as [n] assimilates to the place of articulation of a following sound.

The concept of natural class applies in all the above realisations of the derivative object marker in that in the context of phonemes which share certain features and hence form a natural class the first person object marker [n] changes and acquires some of the features of the phonemes it precedes. For instance, in all these contexts the object marker is syllabic and is realised in different shapes depending on the contexts and the features of the phonemes with which it occurs. The object marker [n] is realised as the bilabial nasal [m] in cases where it is followed by labial sounds such as [p<sup>h</sup>], [p], [b], [f] and [m]; it remains the alveolar nasal [n] when it is followed by alveolars such as [n], [ts<sup>h</sup>], [t], and [tl]; on the other hand it is realised as the nasal [ɲ] before homorganic sounds which are palatal sounds such as [tʃ], [tʃ<sup>h</sup>] and [ɲ]. Further the object marker is realised as the velar nasal [ŋ] when it precedes consonants like [k], [k<sup>h</sup>], [χ], [q<sup>h</sup>] and [h]. On the other hand the object marker is realised as [l] and [r] before [l] and [r], respectively.

In view of this we can posit one underlying form of the first person object marker as the nasal [n] which we can say has different realisations in different contexts as follows:

(7)



It could also be said that the place of the articulation assimilation process which results in these different realisations of the object marker is due to the phonological process of homorganic nasalisation. According to Crystal (1991:167) sounds are said to be homorganic when they are produced at the same place of articulation. On the other hand homorganic nasalisation arises from a juxtaposition of a nasal consonant with some other consonant. This results in the nasal acquiring the same place of articulation with the neighbouring consonant. By the nasal becoming homorganic with the following consonant the first person object marker [n]

assimilates the place of articulation of the initial consonant of the verb to which it is attached.

From the examples in 6(d-n) above, it can be seen that [n] does not change before alveolar sounds. But it changes to the labial [m] before labials, to the palatal [ɲ] before palatals and palato-alveolars and the velar [ŋ] before velars, glottals and uvulars. This can be seen in 6(a-c); 6(o-r) and 6(s-w) above, respectively. The assimilation processes that take place in this regard are formalised in 8(a), 8(b) and 8(c) below. We have formalised three different rules for these processes because of the fact that in each of the rules the output and environment are different:

(8)

a.  $\begin{bmatrix} +\text{nasal} \\ +\text{coronal} \\ +\text{anterior} \end{bmatrix} \rightarrow \begin{bmatrix} +\text{nasal} \\ +\text{labial} \end{bmatrix} / \text{---} \begin{bmatrix} +\text{cons} \\ +\text{labial} \end{bmatrix}$

That is  $[n] \rightarrow [m] / \_\_\_\_\_\_ \text{labial consonants}$ . In the rule in 8(a) the object marker /n/ acquires the feature [+labial] from the following consonants which are labials.

b.  $\begin{bmatrix} +\text{nasal} \\ +\text{coronal} \\ +\text{anterior} \end{bmatrix} \rightarrow \begin{bmatrix} +\text{nasal} \\ +\text{coronal} \\ -\text{anterior} \end{bmatrix} / \_ \_ \_ \begin{bmatrix} +\text{cons} \\ +\text{coronal} \\ -\text{anterior} \end{bmatrix}$

That is [n] → ɲ / \_\_\_\_ palatal consonants. In the rule presented in 8(b), the object marker loses the feature [+anterior] in the presence of [-anterior] sounds.

c. 

+nasal
+coronal
+anterior

 $\rightarrow$ 

+nasal
+back

 / 

+cons
+back

That is  $[n] \rightarrow \eta$  / \_\_\_\_ velar, glottal and uvular consonants. In the rule in 8(c), the object marker /n/ acquires the feature [back] when it occurs before consonants which are [+back]. This is based on one of the redundancy rules we have assumed in this paper which is that [+anterior] = [-back] and vice versa. From all the above examples it can be seen that the derivative syllabic object markers [n], [m], [ŋ] and

[ŋ] form a natural class, they all have the feature [+nasal]. In addition they all become syllabic in the context of homorganic consonants. Moreover they all acquire the place of articulation of the following phonemes to form natural classes with the succeeding phonemes.

From 6(c), (f), (g), (l), (m), (n), (o), (u), (v) and (w) it can be noticed that in some contexts the presence of the first person object marker does not only lead to homorganic nasalisation but also to the strengthening or hardening of some consonants. According to Batibo (2000:170) strengthening or hardening is the process where a sound (usually a consonant) changes in order to become stronger in the presence of nasals or high vowels. He points out the fact that consonants are categorised in terms of their degree of strength as follows:

- (a) Voiceless consonants are stronger than voiced ones.
- (b) Aspirated consonants are stronger than non-aspirated ones.
- (c) Plosives, affricates and nasals are the strongest consonants.
- (d) Fricatives are the next strongest.
- (e) Liquids (trill and laterals) are among the weakest consonants.

From this it can generally be expected that in the process of strengthening the following would occur:

- (9) (a) [+voice] → [-voice]
- (b) [-aspirated] → [+aspirated]
- (c) [+continuant] → [-continuant]
- (d) [+lateral] → [-lateral]

From 6(c), (f), (g), (l), (m), (n), (o), (u), (v) and (w) it can be noticed that the process of strengthening has taken place under the influence of the first person object marker [ŋ] which has triggered the following changes:

(10)

- (a) [b] → [p] i.e. [+voice] → [-voice] in 6(c)
- (b) [d] → [t] i.e. [+voice] → [-voice] in 6(f)
- (c) [s] → [tsʰ] i.e. [-aspirated] → [+aspirated] in 6(g)
- (d) [l] → [t] i.e. [+continuant] → [-continuant] in 6(l)
- (e) [r] → [tʰ] i.e. [+lateral] → [-lateral]) or [+continuant] → [-continuant] in 6(m) and 6(n)

- (f) [ʃ] → [tʃ] i.e. [+continuant] → [-continuant] in 6(o)  
 (g) [χ] → [qʰ] i.e. [+continuant] → [-continuant] 6(u) and 6(v)  
 (h) [h] → [kʰ] i.e. [+continuant] → [-continuant] 6(w)

In the case of 6(c), (o), (u), (v) and (w) there are two phonological rules involved in the syllabification process. In this case the presence of the nasal [n] causes [b], [ʃ], [χ] and [h] to strengthen to [p], [tʃ], [qʰ] and [kʰ], respectively. At the same time [n] undergoes homorganic nasalisation under the influence of the following palatal sound to become [m], [ɲ] and [ŋ]. Our view is that in terms of rule ordering the processes of homorganic nasalisation and strengthening are unordered because even if either was to apply before the other they would not block the application of the other rule because the contexts in which both rules apply are the same. Hence it can be said that in this case the process of strengthening is transparent to the process of homorganic nasalisation and vice versa in Sesotho and Setswana.

## 5.2 Syllabification of Consonants in Word Medial Position

In derivative contexts the nasals [m] and [n] in Sesotho can be syllabic in word medial position due to the attachment of the perfect tense to some verbs. In this position [m] and [n] become syllabic as a result of the attachment of the perfect tense morpheme **-ile** to verbs that have [m] and [n] in the last syllable, respectively. However, in Setswana it is only the nasal [n] which can be syllabified in word medial position due to the attachment of the perfect tense. In this context [m] is not syllabified in Setswana. This is illustrated in (11) below.

(11)

### Sesotho

- |     |                                    |   |                     |
|-----|------------------------------------|---|---------------------|
| (a) | [lɪma] + ile → *lɛmile → *lɛmlɛ    | → | [lɪ.m.mɛ] (lɛmmɛ)   |
|     | 'cultivate'                        |   | 'has cultivated'    |
| (b) | [tlama] + ile → *tlamile → *tlamle | → | [tla.m.mɛ] (tlammɛ) |
|     | 'tie'                              |   | 'has tied'          |
| (c) | [bina] + ile → *binile → *binle    | → | [bi.n.nɛ] (binne)   |
|     | 'sing'                             |   | 'sang'              |
| (d) | [tɪna] + ile → tenile → tenle      | → | [tɪ.n.nɛ] (tenne)   |
|     | 'wear'                             |   | 'wore'              |

**Setswana**

(a)	[nɔna] + ile → *nonile	→ *nonle	→	[nɔ.n.nɛ] (nonne)
	'to be fat'			'has become fat'
(b)	[tɪna] + ile → *tenile	→ *tenle	→	[tɪ.n.nɛ] (tenne)
	'to annoy'			'has annoyed'
(c)	[suna] + ile → *sunile	→ *sunle	→	[su.n.nɛ] (sunne)
	'to kiss'			'has kissed'
(d)	[bina] + ile → *binile	→ *binle	→	[bi.n.nɛ] (binne)
	'to dance'			'has danced'

Syllabification of [m] and [n] word medially which is due to the attachment of the perfect tense morpheme involves two phonological processes. These are deletion and nasalisation. The phonological process of deletion is defined by O'Grady *et al.* (1997) as a process that removes a phoneme from certain phonetic contexts. With regard to the syllabification of consonants in Sesotho and Setswana the phonemes which get deleted are vowels. This is the process of vowel deletion. Initially the affixation of the perfect tense morpheme **-ile** to verbs ending in [m] and [n] results in the deletion of the initial vowel [i] of the perfect tense morpheme as illustrated below:

(12)

- (a) bina → bin + ile → \*binle (Setswana and Sesotho)  
 (b) lema → lem + ile → \*lemle (Sesotho)

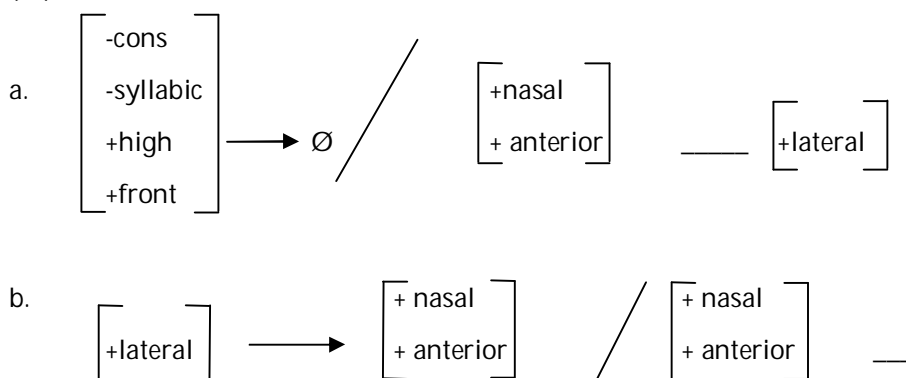
In 12(a) and 12(b) the deletion of [i] results in the forms **\*lemle** and **\*binle** which are unacceptable in the phonotactics of the languages concerned, respectively. To create acceptable forms, after [i] deletion there is the nasalisation of the [l] of the perfect tense morpheme to either [m] or [n]. Hence it could be argued that the process of homorganic nasalisation is necessary because of the phonotactics of Sesotho and Setswana which do not allow the consonant clusters /ml/ (in Sesotho) and /nl/ (in Setswana and Sesotho). In terms of rule ordering it has been observed that the process of deletion triggers the process of the nasalisation of /l/. Deletion precedes and feeds the nasalisation process. According to Crystal (1991:135) a feeding relationship is a situation where the application of one rule creates a structural representation to which another rule can apply or becomes applicable.

This is opposed to the relationship of bleeding which is a situation where a rule removes or destroys the structural representation to which another rule would have applied. Deletion feeds nasalisation in that it creates the structural environment that makes nasalisation necessary.

It can also be said that the case of [l] assimilating to [m] involves manner and place of articulation assimilation. This is because [l] which is a lateral and oral sound becomes a nasal stop due to manner of articulation assimilation. In addition [l] is an alveolar sound which becomes a labial sound by the phonological process of place of articulation assimilation. This is because the preceding sound is a bilabial nasal stop [m]. On the other hand in the case of [l] becoming [n] only manner of articulation assimilation is involved. This is because [l] which is a lateral oral sound changes to a nasal stop [n]. By [l] acquiring some of the features of the preceding phonemes [n] and [m] natural classes are formed.

Both [l] and [n] are alveolar sounds hence the place of articulation does not change. As has been stated above the syllabification of [m] and [n] in the presence of the perfect tense is triggered by the deletion of the initial vowel of the perfect tense morpheme [i]. In fact it is a rule that when the perfect tense morpheme **-ile** is attached to verbs ending in [n] and [m] in Sesotho and to verbs ending in /n/ in Setswana the initial vowel of the perfect tense morpheme is deleted. In turn this leads to the nasalisation of the consonant of the perfect tense [l] to either [n] or [m]. This process of the nasalisation of the perfect tense consonant leads to the syllabification of the last consonant of the verb root. Hence in terms of rule ordering the deletion process feeds the process of syllabification. It creates the necessary condition and environment for nasalisation and syllabification to occur. We have formalised the rules involving [n] and [m] syllabification as follows:

(13)



The rule in 13(a) says that [i] is deleted when it occurs between either [m] or [n] and the lateral [l]. On the other hand the rule in 13(b) says that the lateral [l] becomes

either [m] or [n] when it occurs after either [m] or [n]. The process in 13(a) precedes the process in 13 (b). In turn the process in 13(b) is what leads to the syllabification of [m] and [n].

In word medial position, in derived situations [l] is also sometimes syllabic in Sesotho and Setswana as a result of the attachment of the applicative extension to verbs which end with the sound [l]. This is illustrated in (14) below.

(14)

a.	bala + ela	→	balela	→	[ba.l.la] (balla)
	'read'				'read for'
b.	rapela + ela	→	rapelela	→	[ra.pɛ.l.la] (rapella)
	'pray'				'pray for'
c.	romela + ela	→	romelela	→	[rɔ.mɛ.l.la] (romella)
	'send'				'send to/ for'
d.	bua + ela	→	buelela	→	[bu.ɛ.l.la] (buella)
	'speak/ talk'				'speak for'
e.	tseña + ela	→	tseñelela	→	[tsɛ.nɛ.l.la] (tseñella)
	'enter'				'enter for/go deeper into something/intrude'

On the above examples it should be pointed out that for Setswana some of our respondents stated that the syllabified and non-syllabified forms of the words can be used interchangeably. The view is that the forms of the words where [l] has been syllabified, are dialectal forms which are manifested in rapid speech. However, in Sesotho this process of syllabification is compulsory.

The data analysed has also shown that [l] is also syllabified when the reversive verb extension [olol] is attached to some verbs. The data below illustrates this.

(15)

a.	bofa + olol	→	bofolola	→	[bɔ.fu.l.la] (bofolla)
	'to tie'				'untie'
b.	bipa + olol	→	bipolola	→	[bi.pu.l.la] (bipolla)
	'to cover'				'to uncover/unveil'
c.	epa + olol	→	epolola	→	[ɛ.pu.l.la] (epolla)
	'dig'				'to dig again'



What can be noticed from the examples of the syllabification of [l] with the applicative and reversion extensions is that for syllabification to occur there is initially the deletion of a vowel in the verb extension. In the case of the applicative extension the vowel that is deleted is [e] while the vowel [o] is deleted in the reversion extension. As noted by Doke and Mofokeng (1985), in Sesotho, vowel deletion occurs between two lateral consonants. When this happens the first lateral becomes syllabic.

The deletion of the vowels in both cases results in a situation where we have two l's following one after the other. This results in the syllabification of the first [l] in the sequence. In the examples in (14) the vowel [e] has been deleted between the two laterals whereas in (15) the vowel [o] has been elided between the two laterals as a result of the affixation of the reversion suffix to the verbs. According to distinctive feature phonology it is expected that phonological rules occur in similar environments affecting similar sounds. This seems to be the case with regard to the deletion of [e] and [o]. These phonemes which share the features [-consonantal and +high-mid] are deleted in the same environment. In other words they form the natural class of high-mid vowels. These phonemes get deleted in the same phonetic environment which is between two liquids (laterals). This process is represented in the rule in (16) below.

(16)

$$\left[ \begin{array}{l} \text{-consonantal} \\ \text{+high-mid} \\ \alpha \text{ back} \end{array} \right] \longrightarrow \emptyset \quad / \quad [\text{lateral}] \quad \_\_\_ \quad [\text{lateral}]$$

It would seem as though this deletion occurs as a way of easing articulation, such that one does not have to pronounce two [l]'s successively in consecutive syllables. Moreover two [l]'s in succession is highly unlikely and unnatural in the two languages. However there are cases in which this rule does not apply. For instance, in cases where the syllable which precedes a potential syllabic [l] contains a high vowel [i] or [u], the vowels [e] and [o] are not elided. This can be noticed in the following words.

- (17)
- |    |             |               |
|----|-------------|---------------|
| a. | bu.lɛ.la    | 'open for'    |
| b. | ko.tu.lɛ.la | 'harvest for' |
| c. | du.lɛ.la    | 'wait for'    |
| d. | di.lɛ.la    | 'smear for'   |
| e. | si.lɛ.la    | 'grind for'   |

Our view is that high vowels seem to block or bleed the syllabification of [l] in these contexts. They destroy the environment in which syllabification could have occurred.

We have also noticed that there are some instances where the vowels do not get deleted when they appear between two laterals and the vowel in the preceding syllable of the word is not [+high]. This can be noticed in the Sesotho words below.

- (18)
- |    |             |                |
|----|-------------|----------------|
| a. | po.lɛ.lo    | 'sentence'     |
| b. | ha.la.lɛ.la | 'is holly'     |
| c. | ho.po.lɛ.la | 'remember for' |

These instances can be regarded as exceptions to the rule of deletion. As noted by Kolorenc (2007) phonological rules do not cover all words. There are always some contexts in which they cannot apply.

### 5.3 Syllabification of Consonants in Word Final Position

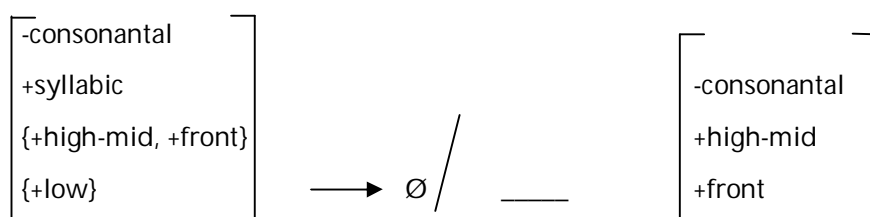
Among the derivative syllabic consonants under discussion it is only the nasal [ŋ] that can occur in word final position. It appears as a syllabic nasal word finally as a locative suffix **-eng**, which is added to nouns to form the locative forms as in the following:

- (19)
- |    |                 |              |   |              |                      |
|----|-----------------|--------------|---|--------------|----------------------|
| a. | thaba [thaba]   | thaba + eng  | → | [tʰa.be.ŋ]   | thabeng              |
|    | 'mountain'      |              |   |              | 'on/at the mountain' |
| b. | tsela [tsɪla]   | tsela + eng  | → | [tsɪ.le.ŋ]   | tseleng              |
|    | 'road'          |              |   |              | 'on/at the road'     |
| c. | sekolo [sɪkolo] | sekolo + eng | → | [sɪ.ko.lo.ŋ] | sekoleng             |
|    | 'school'        |              |   |              | 'at/in school'       |
| d. | kereke [kereke] | kereke + eng | → | [kerekeŋ]    | kerekeng             |
|    | 'church'        |              |   |              | to/at church         |
| e. | sekoti [sekoti] | sekoti + eng | → | [sekotiŋ]    | sekoting             |
|    | 'ditch'         |              |   |              | 'in/at the ditch'    |
| f. | mobu [mobu]     | mobu + eng   | → | [mobuŋ]      | mobung               |
|    | 'soil'          |              |   |              | 'to/ at the soil'    |

- |    |                          |             |   |                                 |         |
|----|--------------------------|-------------|---|---------------------------------|---------|
| g. | huku [huku]              | huku + eng  | → | [hukuŋ]                         | hukung  |
|    | 'corner'                 |             |   | 'at the corner'                 |         |
| h. | ofisi [ofisi]            | ofisi + eng | → | [ofisiŋ]                        | ofising |
|    | 'office'                 |             |   | 'at the office'                 |         |
| i. | mosu[musu]               | mosu + eng  | → | [mosuŋ]                         | mosung  |
|    | 'thorny indigenous tree' |             |   | 'at the thorny indigenous tree' |         |

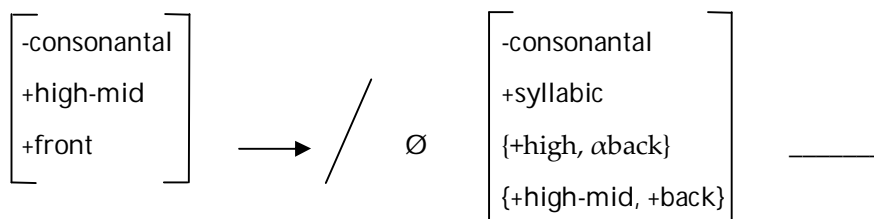
This rule applies to all situations where the locative suffix *-eng* is attached as a suffix to nominals to form locative forms. It has also been observed that the addition of the locative suffix *-eng* to nominal forms induces the deletion of either the final vowel of the nominal stem or the initial vowel of the locative *-eng*. In the examples above the final vowels of the nominal stems ending in [a] and [ɛ] are deleted before the locative suffix initial vowel [e]. On the other hand when the nominal form ends in the vowels [o], [i] and [u] the initial vowel [e] of the locative suffix *-eng* is deleted. We present the two rules reflecting these changes below:

(20)



This means that the low vowel [a] and the high-mid front vowel [e] are deleted in the context before a high-mid front vowel [e]. The second rule involved in the examples in 19 is as follows:

(21)



This rule means that the high-mid front vowel [e] is deleted in the context after the high back or front vowels [u] or [i] and in the context of the high-mid back vowel

[o]. The manifestation of the rule in (20) is illustrated in 19(a), (b) and (d) above. On the other hand the rule in (21) is manifested in 19(c), (e), (f), (g), (h) and (i) above.

As has been stated above, syllabic /r/ is only found in Setswana. It has also been established that among the syllabic consonants under discussion it is only /r/ which does not occur in derivative contexts.

## **6. CONCLUSION**

This chapter has discussed the occurrence of the derivative syllabic consonants [m], [n], [ɲ], [ŋ] and [l] in Sesotho and Setswana. The focus of the discussion has been the description of the contexts in which these syllabic consonants occur and the phonological processes that are involved in making the consonants syllabic. The process of the syllabification of consonants in Sesotho and Setswana has been recognised in some previous literature in the languages (cf. Cole (1955), Kunene (1961), Guma (1971) and Doke and Mofekeng (1985)).

As it has been pointed out above these scholars mainly attribute the process of syllabification to the phonological process of the deletion of vowel phonemes between two [l]'s which results in one of the [l]'s in this context becoming syllabic. The conclusions of these scholars seem to have been based on an analysis of limited contexts of the occurrence of derivative syllabic consonants in the languages concerned. Their conclusions seem to have been based on the process as it applies to the formation of reversive verbs and processes of perfect tense marking and applicativisation.

Our analysis has shown that this indeed is the case in the syllabification process that affects [l] with perfect tense marking and applicativisation. However, using data from various other contexts in which derivative syllabic consonants are manifested in the languages our analysis has further shown that there are other contexts of derivative syllabic consonants in the languages whose existence can be attributed to various other phonological processes not emphasised on before.

With specific reference to data involving formation of reversive verbs, object marking, perfect tense marking, applicativisation and locativisation of some word forms we have shown that Sesotho and Setswana have other derivative syllabic consonants, namely [n], [m], [ɲ], [ŋ] and [l] which occur before homorganic consonants. In other words, syllabic [m] occurs before bilabial sounds, [n] before alveolar sounds, [ɲ] appears before palatal and palato-alveolar sounds, [ŋ] appears before velar, glottal and uvular sounds and [l] occurs before another [l]. As for positions in which syllabic consonants occur, it has been stated that the syllabic consonants can occur word initially and medially. The syllabic consonant [ɲ] is the only one that can occur word finally.

In relation to the phonological processes involved in the syllabification of the consonants it is concluded that these are deletion, place of articulation assimilation (or what we have also referred to as homorganic nasalisation) and manner of articulation assimilation.

It has also been noted that the syllabification of [m] and [n] involving the affixation of the perfect tense to verbs ending in [m] and [n] involves two phonological processes which have a feeding relationship. In this case the initial vowel of the perfect tense morpheme **-ile**, [i] is deleted and in turn this triggers the homorganic nasalisation of the [l] of the perfect tense affix to the preceding nasal.

In the syllabification of [l] when the applicative affix, **-ela**, is attached to some verbs ending in the lateral [l] there is also the deletion of the initial vowel of the applicative extension [e] which leads to the syllabification of the [l] immediately preceding the applicative extension. It has also been established that [l] can be syllabified in some instances with the attachment of the reversive extension by the same process that applies to syllabification with perfect tense marking and applicativisation. The process involves deletion of the second [o] of the reversive extension which in turn triggers the syllabification of [l] in the context in which the process of deletion has taken place.

Among the derivative syllabic nasals discussed it is only the nasal [ŋ] that can occur in word final position. It appears as a syllabic nasal word finally as a locative suffix **'-eng'**, which is added to nouns to form the locative forms. In our discussion it has been shown that the addition of the locative suffix **-eng** to nominals induces the deletion of the last vowel of either the nominal or the initial vowel of the locative suffix.

## REFERENCES

- Atkinson, M. et al. 1982. *Foundations of General Linguistics*. London: Unwin Hyman Ltd.
- Batibo, H. 2000. System of the Sounds of Africa. *African Voices: An Introduction to the Languages and Linguistics in Africa*, edited by V. Webb & Kembo-Sure. Cape Town: Oxford University Press of Southern Africa.160-196.
- Bird, S. 1997. Dschang Syllable Structure. <http://cogprints.org/2183/00/syllable.pdf>
- Cairns, C. & M. Feinstein. 1982. Markedness and the Theory of the Syllable. *Linguistic Inquiry* 13:193-226.
- Chomsky, N. & M. Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.
- Cole, Desmond. 1955. *An Introduction to Tswana Grammar*. Cape Town: Longman
- Crystal, D. 1991. *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell Publishers.
- Demuth, K. 1993. Issues in the Acquisition of the Sesotho Tonal System. *Journal of Child Language* 20:275-301.
- Demuth, K. Accessed in 2007. Sesotho Speech Acquisition. <http://www.cog.brown.edu/People/demuth/articles/sm%20-%20Demuth%20-%20Part%20II%20Sesotho-3.pdf>
- Doke, G.M. & S.M. Mofokeng. 1985. *Textbook of Southern Sotho Grammar*. Cape Town: Maskew Miller Longman.
- Guma, S.M. 1971. *An Outline Structure of Southern Sotho*. Pietermaritzburg: Shuter & Shooter.
- Guthrie, M. 1948. *The Classification of Bantu Languages*. London: International African Institute.
- Hyman, L.M. 1975. *Phonology: Theory and Analysis*. New York: Holt, Rinehart and Winston.
- Jakobson, R., G. Fant & M. Halle. 1952. *Preliminaries in Phonological Analysis*. Cambridge, MA: MIT.
- Kolorenc, J. Accessed in 2007. Evolving Phonological Rules Using Grammatical Evolution. <http://itakura.kes.vslib.cz/kes/public/post04k>.
- Kunene, D. P. 1961. *The Sound System of Southern Sotho*. PhD Thesis. Cape Town: University of Cape Town.

Martin, A. Accessed in 2003. Postnasal Vowel Deletion in Navajo.

[http://www.linguistics.ucla.edu/people/grads/amartin/Navajo\\_vowel\\_deletion.pdf](http://www.linguistics.ucla.edu/people/grads/amartin/Navajo_vowel_deletion.pdf)

O'Grady, W., M. Dobrovolsky & F. Katamba. 1997. *Contemporary Linguistics: An Introduction*. London: Longman.

## CHAPTER 6

### **PHONETIC DATA AND PHONOLOGICAL THEORY: A REPORT FROM THE CIVILI VOWEL DURATION ISSUE**

Hugues Steve Ndinga-Koumba-Binza

*Centre for Text Technology, North-West University, Potchefstroom, South Africa  
Langue, Culture et Cognition (LCC), Université Omar Bongo, Libreville, Gabon  
22602569@nwu.ac.za*

#### **1. INTRODUCTION**

This chapter intends to report on the constitution and the nature of data that was used in the study of the Civili vowel duration issue. The term “data” is understood as utterances, which form the basis for linguistic investigation. The traditional conception of linguistic data is limited to the observable patterns of speech and writing, especially when recorded and gathered together within a corpus. The interaction between data and theories has recently put this definition to assessment in different linguistic sub-fields. In fact, accurate descriptions and adequate theories require reliable data.

The paper aims to contribute to the definition of the nature of efficient data for a systematic and reliable phonological study within the framework of the so-called *Phonetics-Phonology Interface Debate* (P-PID). It looks at the implication of data within the relationship between phonetics and phonology with reference to the description of vowel duration in the Civili (H12a) language.

This chapter mainly focuses on speech and acoustic data. An account of how perceptual data should be acquired and how perceptual data were used in Ndinga-Koumba-Binza (2008) has indeed been the subject of other publications by Ndinga-Koumba-Binza (2009, 2011 & 2012), Ndinga-Koumba-Binza and Roux (2009a & 2009b), and Roux and Ndinga-Koumba-Binza (2011).

#### **2. PREVIOUS STUDIES ON CIVILI VOWEL DURATION: A DATA ISSUE**

Vowel duration has been an issue in the phonological system of Civili. Describing the issue, Ndinga-Koumba-Binza (2004) has noted that the so-called Civili vowel length desperately needs to be re-examined because previous works on the phenomenon in the language (Marichelle 1902, Ndamba 1977, Blanchon 1990, Mabika Mbokou 1999, Ndinga-Koumba-Binza 2000) referred to impressionistic phonetic observations and do not rely on experimentally-gathered and verified



data. This view launched the research work by Ndinga-Koumba-Binza (2008 & 2012), which is an experimental phonetic investigation into the vowel duration of Civili. The argumentation in this paper comes within the Phonetics-Phonology Interface Debate (P-PID), which is known as a specific focus on the relationship between the phonological component and the phonetic component. This debate has particularly shown that:

- (i) phonetic data can confirm or disprove phonological analyses (Garnes 1973), and
- (ii) phonetic output, as predicted by phonological theories, could lend credibility to or negate the analysis (Ohala 1990).

Ndinga-Koumba-Binza (2004:197-198, 2008:29-53, 2012:21-41) has assessed existing data stemming from studies by Marichelle (1902), Ndamba (1977), Blanchon (1990), Mabika Mbokou (1999), and Ndinga-Koumba-Binza (2000). These data have been found impressionistic and inadequate for a reliable processing of Civili vocalic quantity. They could not offer satisfactory analyses for such a specific requirement as vowel lengthening in any language. In fact, as Roux (1995:197) points out,

linguists working within African languages have up to this day been quite complacent to rely almost exclusively on the impressionistic judgments of a 'trained phonetician' in compiling primary data.

Data in studies by Marichelle (1902), Ndamba (1977), Blanchon (1990), Mabika Mbokou (1999), and Ndinga-Koumba-Binza (2000) were also impressionistic, as they were gathered from general wordlists made to serve in any kind of linguistic study. It may be assumed that these data were gathered not only for phonetic analyses, but also for morphosyntactical (Ndamba 1977), analogical (Mabika Mbokou 1999), phonological and tonological (Blanchon 1984, 1990; Blanchon & Nsuka-Nkutsi 1984; Ndinga-Koumba-Binza 2000) descriptions, as well as for a dictionary compilation purpose (Marichelle 1902). Also note that these studies did not have any acoustic data as primary source.

The works preceding Ndinga-Koumba-Binza (2000) could not give satisfactory answers to the issue of vowel lengthening in Civili, amongst other reasons because the phonetic data of these previous studies on the Civili vowel-sound system are very limited in nature. They are limited because they were gathered randomly without any expectation of further studies, apart from phonetic-articulatory description that enables phonological procedures such as segments inventory, features distinction and structures study.

### 3. TYPES OF DATA

In different linguistic sub-fields, the interaction between data and theories has recently put to assessment the traditional definition of the concept “data”. An “efficient way” of dealing with data is to evaluate its reliability by statistical techniques (cf. Tams 1999:2). The author of this chapter advocates for three types of data to be collected for a proper study of the Civili vowel duration, i.e. (i) speech (linguistic) data, (ii) acoustic data and (iii) perceptual data.

It is important to point out that perceptual data are equally significant and reliable in a phonetic and phonological study such as vowel duration in Civili. There is a distinction between the physical (acoustic) duration and the perceived duration. This is why Malmberg (1974:191) had to distinguish the physical duration from distinctive length. Indeed, in the process of human language communication there is a difference between the produced signal (acoustics and/or articulatory) and the perceived signal (perception). Nevertheless, there is an increasing interest of studying the production (acoustic and/or articulatory) and perception in parallel in order to find out whether perception is based on the production (articulatory or acoustic) templates and in which manner production is controlled by auditory templates (cf. Eerola *et al.* 2003; Ru *et al.* 2003; Frieda, Walley *et al.* 2000). For the specific case of Civili as presented in Ndinga-Koumba-Binza (2008), perceptual data gathered through acoustic stimuli and perceptual tests were analyzed.

As for speech data, it is also herein believed that this “efficient way” should start with an efficient procedure for speech data acquisition. If speech data are not acquired properly, this will definitely affect the results of the analyses. For the specific case of Civili vowel duration, it is first suggested that a reliable wordlist should be built. The wordlist should specifically be aimed at getting data for measurements and experiments for a vowel duration study. Textbooks in phonetics – such as Laver (1994), Ashby (1995), Clark and Yallop (1995), Kent and Read (2002) and Collins and Mees (2003) – show a common agreement that the primary goal of phonetic research is threefold. Bird and Gick (2006:463) outline this goal as:

- (i) to document the different sounds that occur in natural languages (e.g. Ladefoged & Maddieson 1996);
- (ii) to understand the acoustic and articulatory properties of these sounds (e.g. Miller-Ockhuizen 2003); and
- (iii) to evaluate experimentally theories and models of phonetic and phonological structure (e.g. Bird & Caldecott 2004).

The study of Civili vowel duration came within the framework of the latter aspect (iii), i.e. the experimental evaluation of theories and models of phonetic and

phonological structure. It was, however, still crucial to acquire speech data from native speakers of the studied language in order to adhere to the following guideline by Bird and Gick (2006:463):

In some cases, speakers can be recorded in a laboratory setting; this is practical with many languages spoken in urban areas. When speakers cannot be brought to a laboratory, however, it is necessary to conduct phonetic fieldwork, i.e., to record speech outside of a laboratory setting.

In this regard, Ndinga-Koumba-Binza (2008 & 2012) applied the above procedure capturing speech data using the computer software PRAAT (for details cf. [www.praat.org](http://www.praat.org)). Speech data for the present study were acquired “*outside a laboratory setting*” (Bird and Gick 2006:463). The corpus, on which the study was based, was made up according to general principles that governed the compilation of a set of words compiled for the data acquisition. Given the need for new data for the experimental study of the problematic phonological phenomenon in Civili, a representative corpus for such an experimental study of the Civili vowel duration was constructed. The purpose of the corpus was to collect Civili speech data. Its function was to enable one to perform acoustic analysis and vowel duration measurements. Linguists have often defined the term “corpus” as “*a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description*” (Kennedy 1998:1). However, for the specific purpose of this study, a corpus was taken to be a sample of language that has been collected in order to “*provide an empirical basis for describing and mapping out the use of language systems*” (De Klerk 2002:25). This was a smaller purpose-designed corpus intended for a phonetic-phonological study. Like many phonetic studies (e.g., Ohala 1997:686), the aim was to use a set of minimal pairs in order to ascertain which of the cues differentiate between phonologically-specified contrasts. However, distinctive contrasts between long and short vowels are particularly rare in Civili. Therefore, the corpus consisted of:

- (i) a range of words containing long-sounding and short-sounding vowels, on the one hand; and
- (ii) a range of sentences and phrases containing the same words in both syntactical positions of subject and object, on the other hand.

The words included into the corpus were based on the following previous studies on the Civili language spoken in Mayumba: Blanchon (1984 and 1990); Mabika Mbokou (1999); and Ndinga-Koumba-Binza (2000). In general, corpora have to meet certain criteria, which are usually subsumed under the general principle of representativeness. Kučera (2002:246) indicates that,

in smaller specialized corpora..., the representativeness of the corpus may be a relatively straightforward principle: often it

simply means the inclusion of all the relevant texts in their authentic form.

According to Kučera (2002:146), the principle of representativeness has been

used and referred to rather loosely and vaguely in both corpus and non-corpus linguistics, and the differences between existing corpora suggest that there are differing views on how the general concept translates into the size and structure of a large versatile corpus.

Compared to various corpora as dealt with in the specific field of corpus linguistics, the corpus of the Civili study was a very limited and focused corpus. The reason for this was that, due to the aim of obtaining new and specific data for the measurement and analysis of vowel duration in Civili, it was necessary to build a reliable and specific corpus to be used for data recording. The corpus contains 384 entries of single words in isolation and 768 sentences and phrases. Table 3 below contains a sample of the Civili wordlist for a Vowel Duration Study. The first column shows words in Civili. The English translation of these words is presented in the second column in italics. The third column shows the French equivalents of the words.

**Table 1:** Sample of the Civili wordlist for a Vowel Duration Study  
(cf. Ndinga-Koumba-Binza 2012:48)

<b>/i/</b>		
<b>ciima</b>	<i>thing</i>	chose
<b>lizina</b>	<i>name</i>	nom
<b>/e/</b>		
<b>lisefu</b>	<i>smile</i>	sourire
<b>siseenda</b>	<i>thorns, prickles</i>	épines
<b>/a/</b>		
<b>lisaafi</b>	<i>lung</i>	poumon
<b>ndaka</b>	<i>tongue</i>	langue
<b>/o/</b>		
<b>n'cyoodu</b>	<i>sword</i>	épée
<b>tusoku</b>	<i>spear</i>	lance
<b>/u/</b>		
<b>n'kuumba</b>	<i>navel</i>	nombril
<b>kutu</b>	<i>ear</i>	oreille

The corpus also meant to highlight some factors that often influence vowel duration. Among such factors are, according to Kent and Read 2002:127 and Klatt 1976), tense-lax (long-short) feature of the vowel, vowel height, syllable stress, speaking rate, voicing of a preceding or following consonant, place of articulation of a preceding or following consonant; and various syntactic factors, such as utterance position. According to Kent and Read (2002:127),

some of these are inherent durational attributes (e.g. tenseness or laxness, vowel height), and other are determined by the suprasegmental properties or phonetic context (e.g., stress, speaking rate, consonant environment).

Since the present study built on previous phonological studies of the vowel system of Civili, it was particularly concerned with the following factors:

- (i) The position of the word to which the studied vowel belongs, i.e.,
  - a. the word in isolation;
  - b. the word as object in a phrase or sentence; and
  - c. the word as subject in a phrase or sentence.
- (ii) the context or environment of the vowel (immediate or direct phonological context, segments before and segments after).

According to Mabika Mbokou (1999:31-32) and Ndinga-Koumba-Binza (2000:76-77, 79-85), duration variation could occur in the following environments:

- (i) /C\_NC/ (a vowel is long when it is followed by a nasal cluster).
- (ii) /CG\_C/ (a vowel is long when it follows a consonant-glide sequence).

In fact, a number of phonological descriptions and analyses (Batibo 1985:23; Clements 1986:45; Odden and Odden 1999:2; Myers Hansen 2005:317) have stated that any vowel is lengthened or is long when,

- (i) it precedes a nasal segment, leading to the following formalism:

(17) V       $\longrightarrow$     [+long] /      \_\_N

- (ii) it follows a consonant-glide segment, thus leading to the following formalism:

(18) V       $\longrightarrow$     [+long] /      CG\_\_

These claims apply to Civili, as seen in Mabika Mbokou (1999:31-32) and Ndinga-Koumba-Binza (2000:76-77, 79-85). Nevertheless, in order to broaden theoretical formalised contexts above, the organisation of the corpus also considered the inclusion of the following contexts:

- (i) /C\_N/ (when the vowel is followed by a single nasal consonant).
- (ii) /C\_C/ (when the vowel precedes a consonant stop).
- (iii) /C\_L/ (when the vowel precedes a liquid consonant).

Once appropriate speech data were acquired, the researcher had then to proceed with acquiring acoustic data. Gathering acoustic data consists of conducting a phonetic investigation into the acoustic properties of recorded speech (Fujimura & Erickson 1997; Ladefoged 1997, Laver 1994). Acoustic data are necessarily acquired within an acoustic experiment. The acoustic experiment might simply state that speech data should acoustically be processed. In the study of the Civili vowel duration, acoustic data were acquired within a process called speech data analysis. Speech analysis is usually defined in terms of the following three components (Ifeachor & Jervis 1993; Robinson 1998):

- (i) Analysis of speech sounds, taking into consideration their method of production.
- (ii) The extraction of "interesting" information as an acoustic vector.
- (iii) The level or processing between the digitised acoustic waveform and the acoustic feature vector.

The Civili vowel duration study was concerned with the second component. It is important to note that an acoustic vector is a representation of the speech sound at a specific time period of production of that speech sound. For example:

- (i) the short-term power spectra;
- (ii) a representation of the vocal tract shape; and
- (iii) an estimation of the formant frequencies and bandwidths.

Note that in the field of phonetic sciences duration is viewed as one of the acoustic characteristics of phonic units such as consonants and vowels (Klatt 1976; Erickson 2000; Kent & Read 2002). Note that this is not an acoustic study of Civili vowels per se, as this was not one of the aims of the present study. For the specific purpose of the Civili vowel duration study, the core of the acoustic analysis consisted of labelling and segmentation and vowel duration measurements.

It sought to distinguish vowels (short, long and/or double) according to the feature of duration. Therefore, detailed analyses were made of vowel segments of Civili, with specific attention given to the status of long and/or double vowels. Jones (2001:11) mentioned that most acoustic studies on ambiguous sentences *"have revealed unanimity in their results, indicating that the fundamental frequency (F0) and segmental duration are important factors for disambiguating syntactically ambiguous sentences"*. It was hypothesised that this same factor, namely segmental duration,

may also be important for disambiguating Civili words within minimal pairs. The study endorsed the importance of segmental duration (or natural duration) in order to differentiate between short vowels, long or lengthened vowels and sequencing of identical vowels with regard to the issue of vowel duration.

In general, such an analysis attempts to identify acoustic differences between original vocalic statements. This is in agreement with Jones (2001:12) who states that the results should then be *“also statistically analyzed and then perceptually tested to increase the validity and authenticity of the research output”*. PRAAT, the speech editing tool, has functions for speech analysis, speech synthesis, learning algorithms, labelling and segmentation, speech manipulation, listening experiments, etc. (cf. Boersma & Weenink 1992-2001). The study, however, specifically made use of PRAAT’s functions for speech analysis at primary level, labelling and segmentation, and listening experiments. The purpose of labelling and segmentation was to identify the relevant segments, i.e., identified vowels. Labelling and segmentation in PRAAT include:

- (i) annotation, i.e., labelling intervals/waveforms and time points on multiple tiers;
- (ii) the use of the phonetic alphabet; and
- (iii) the use of sound files up to 2 gigabytes (3 hours), i.e., LongSound.

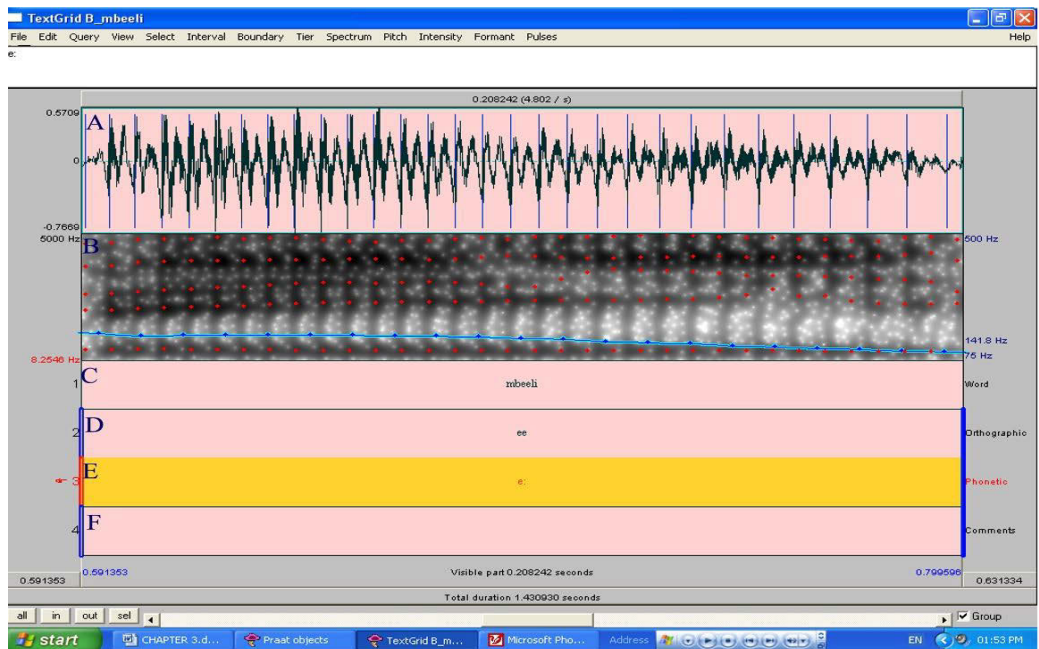
Data are labelled in a TextGrid object, which is one of the types of objects in PRAAT. Items of the wordlist were segmented and labelled from recorded utterances in the corpus, such that each phoneme in the transcription was aligned with its corresponding sound event. Time alignment for each word as a whole was provided. In the context of sentences and phrases, the segmental transcription and time alignment included only the test word; no segments from adjacent words were included in the transcription. Thus, if a segment or a pause is seen in the transcription, it was part of the pronunciation of that token. Both beginning and end of time points of selected words were marked, and a segmental phonetic transcription for each word was also provided.

In the case of mid-vowels, a distinction is made at the phonetic level where these vowels are transcribed as [e] and [ɛ] for the [+FRONT] ones, and as [o] and [ɔ] for the [+BACK] ones. At the orthographic level, they are represented by /e/ and /o/, as respectively retained in the orthography of Gabonese languages (Idiata 2002) and in most of phonological descriptions of Civili (see Ndinga-Koumba-Binza 2006 for details on the quality of mid-vowel in Civili).

A transcription may be narrow or broader. It may be a transcription of how a word is generally uttered in a particular language, or it may attempt to capture the actual variation in how a word is uttered by one person on one particular occasion. Which

level the researcher chooses to work at depends on the aim of the research. Figure 1 below, which is a PRAAT screenshot, displays the following: **A** indicates the source waveform, of recording of the word *mbeeli*; **B** is the spectrogram associated with the waveform; **C** is the tier for the name of the structure – the word *mbeeli* in this case; in **D** the phoneme under investigation is noted – the long-sounding vowel [e:] of *mbeeli*; in **E** the specific phonetic realisation of the phoneme indicated in D is mentioned; and **F** indicates the space left for any relevant comments.

**Figure 1:** A typical PRAAT screenshot showing the /ee/ within the word /mbeeli/  
(cf. Ndinga-Koumba-Binza 2012:56)



Tags and transcriptions were inserted manually. The methodology of labelling and segmentation was both visual – using the source waveform (**A** in Figure 1) and the spectrogram (**B** in the figure) associated with the waveform – and auditory, as it also entailed listening to the marked speech segments.



#### 4. CONCLUSION

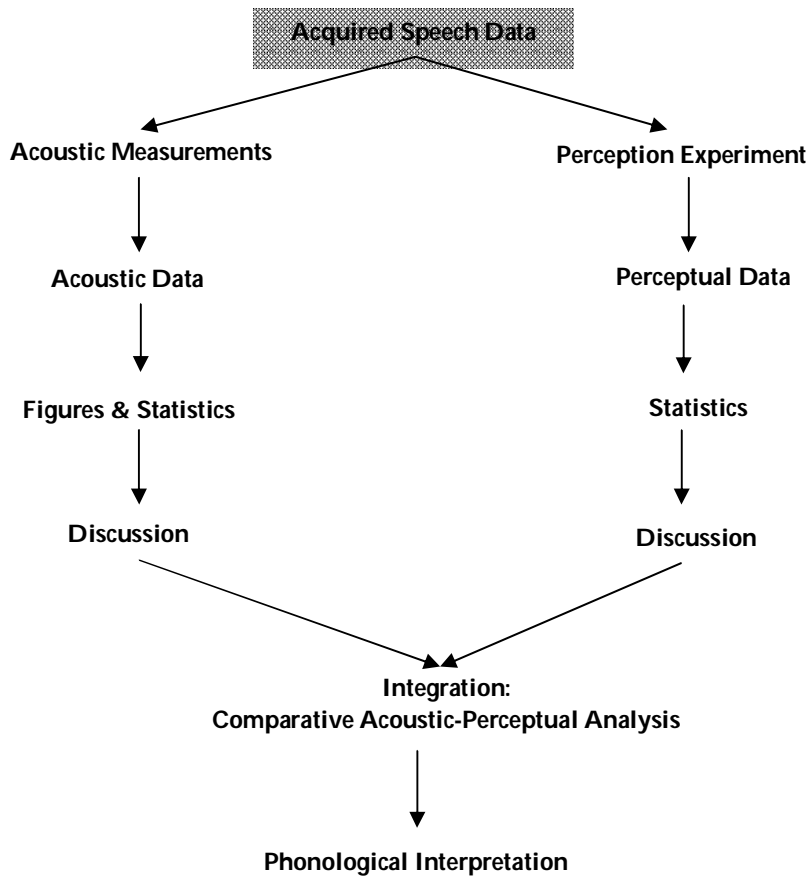
The current description of vowel duration in Civili incorporates statistically analyzed acoustic information together with experimentally verified perceptual information (cf. Ndinga-Koumba-Binza & Roux 2009). This paper does not only contribute to the phonetics-phonology interface debate, but mainly to the domain of data acquisition.

It yields support for the incorporation of phonetic representation into the phonological description. This is referred to as the integration of both domains. It gives support to the idea that experimental data can enhance phonological analyses. It has presented a procedure of gathering data for a specific study of vowel duration. This consists on making observations in order to make reliable statements.

Deriving phonetic facts from data should be the whole point of an experimental procedure which should consist of gathering appropriate data, evaluating and analyzing the data by statistical methods, arriving at a generalisation concerning the data, expressing this generalisation as a fact, as well as fitting the facts into a model. The experimental procedure as conducted for the study of the Civili vowel duration in this paper is summarised in Figure 2 below.

Acquired acoustic and perceptual data should be interpreted in terms of current non-linear phonological models in quest for a credible analysis. It often happens that if the phonological theory is not able to account for data the theory is changed. This agrees with Garnes (1973:273) who states that "*[i]f the resulting surface forms are not attested in the spoken language itself, the phonological analysis loses its credibility*".

**Figure 2:** *Experimental Procedure for Data Acquisition & Processing*  
(cf. Ndinga-Koumba-Binza 2012:133)



## REFERENCES

- Ashby, P. 1995. *Speech sounds*. London & New York: Routledge.
- Batibo, H. 1985. *Le kesukuma (langue bantu de Tanzanie): phonologie, morphologie*. Paris: Editions Recherche sur les Civilisations.
- Bird, S. & M. Caldecott. 2004. Timing differences in St'át'imcets glottalised resonants: linguistic or biomechanical? *Proceedings of the 10<sup>th</sup> Australian International Conference on Speech Science & Technology Association (STT)*. Sydney: Macquarie University/Australian Speech Science & Technology Association Inc.328-333.
- Bird, S. & B. Gick. 2006. Phonetics: field methods. *The encyclopedia of language and linguistics*, edited by K. Brown. Volume 9:463-467. Oxford: Elsevier.
- Blanchon, J.A. 1984. Présentation du yi-lumbu dans ses rapports avec le yi-punu et le ci-vili à travers un conte traditionnel. *Pholia* 1:7-35. Reprinted in Blanchon, J.A. 1999. *Douze études sur les langues du Gabon et du Congo-Brazzaville*. München: Lincom Europa.5-31.
- Blanchon, J.A. 1990. Civili. *Revue Gabonaise des Sciences de l'Homme* 2:141-142.
- Blanchon, J.A. & F. Nsuka-Nkutsi. 1984. Détermination des classes tonales des nominaux en ci-vili, en i-sangu et en i-nzebi. *Pholia* 1:37-45.
- Boersma, P. & D. Weenink. 1992-2001. *Praat: A system for doing phonetics by computer*. <http://www.praat.org> Accessed: 22-04-2007.
- Clark, J. & C. Yallop. 1995. *An introduction to phonetics and phonology*. Oxford: Blackwell Publishers.
- Clements, G.N. 1986. Compensatory lengthening and consonant gemination in LuGanda. *Studies in compensatory lengthening*, edited by L. Wetzels and E. Sezer. Dordrecht: Foris.37-77.
- Collins, B. & I.M. Mees. 2003. *Practical phonetics and phonology: A resource book for students*. London and New York: Routledge.
- De Klerk, V. 2002. Towards a corpus of Black South African English. *Southern African Linguistics and Applied Language Studies* 20:25-35.
- Eerola, O., J.P. Laaksonen, J. Savela & O. Aaltonen. 2003. Perception and production of the short and long Finnish [i] vowels: Individuals seem to have different perceptual and articulatory templates. *Proceedings of the 15<sup>th</sup> International Congress of Phonetic Sciences*, edited by M.J. Solé, D. Recasens & J. Romero. Barcelona: Casual Productions.989-992.

- Erickson, M.L. 2000. Simultaneous effects on vowel duration in American English: A covariance structure modeling approach. *The Journal of the Acoustical Society of America* 108:2980-2995.
- Frieda, E.M., A.C. Walley, J.E. Flege & M.E. Sloane. 2000. Adults' perception and production of the English vowel /i/. *Journal of Speech, Language and Hearing* 43: 129-143.
- Fujimura, O. & D. Erickson. 1997. Acoustic phonetics. *The handbook of phonetic sciences*, edited by W.J. Hardcastle & J. Laver. Oxford/Cambridge: Blackwell Publishers.65-115.
- Garnes, S. 1973. Phonetic evidence supporting a phonological analysis. *Journal of Phonetics* 1:273-283.
- Idiata, D.F. 2002. *Il était une fois les langues gabonaises*. Libreville: Editions Raponda-Walker.
- Ifeachor, E.C & B.W. Jervis. 1993. *Digital speech processing: A practical approach*. Boston, MA: Addison-Wesley.
- Jones, C.J.J. 2001. *Queclaratives in Xhosa: An acoustic and perceptual analysis*. Unpublished Doctoral Dissertation. Stellenbosch University.
- Kennedy, G. 1998. *An Introduction to corpus linguistics*. London and New York: Longman.
- Kent, R.D. & C. Read. 2002. *The acoustic analysis of speech*. Albany, NY: Singular Thomson Learning.
- Klatt, D.H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* 59:1208-1221.
- Kučera, K. 2002. The Czech national corpus: Principles, design, and results. *Literary & Linguistic Computing* 17(2):245-257.
- Ladefoged, P. 1997. Instrumental techniques for linguistic phonetic fieldwork. *The handbook of phonetic sciences*, edited by W.J. Hardcastle & J. Laver. Oxford/Cambridge: Blackwell Publishers.137-166.
- Ladefoged, P. & I. Maddieson. 1996. *The sounds of the world's languages*. Oxford: Blackwell Publishers.
- Laver, J. 1994. *Principles of phonetics*. Cambridge: Cambridge University Press.
- Mabika Mbokou, L. 1999. *Les phénomènes analogiques en civili: étude phonologique et morphologique*. Unpublished Master's Thesis. Libreville: Université Omar Bongo.

- Malmberg, B. 1974. *Manuel de phonétique générale*. Paris: Editions Picard. Collection Connaissance des Langues.
- Marichelle, C. 1902. *Dictionnaire vili-français*. Loango: Imprimerie de la Mission.
- Miller-Ockhuizen, A. 2003. *The phonetics and phonology of gutturals: A case study from Ju'hoanisi*. New York: Routledge.
- Myers, S. & B.B. Hansen. 2005. The origin of vowel-length neutralisation in vocoid sequences: evidence from Finnish speakers. *Phonology* 22 (2005):317-344.
- Ndamba, J. 1977. *Syntagme nominal et groupe nominal en vili (H12): Langue bantoue du Congo*. Unpublished Doctoral Dissertation. Paris: Université Sorbonne Nouvelle.
- Ndinga-Koumba-Binza, H.S. 2000. *Phonologie du civil de Mayumba: Langue bantu du Gabon (H12a)*. Unpublished Master's Thesis. Libreville: Université Omar Bongo.
- Ndinga-Koumba-Binza, H.S. 2004. Vowel duration issue in Civili. *South African Journal of African Languages* 24(3):189-201.
- Ndinga-Koumba-Binza, H.S. 2006. Mid-vowels and vowel harmony in Civili. *South African Journal of African Languages* 26(1):26-39.
- Ndinga-Koumba-Binza, H.S. 2008. *Phonetic and phonological aspects of the Civili vowel duration: An experimental approach*. Doctoral dissertation. Stellenbosch: Stellenbosch University.
- Ndinga-Koumba-Binza, H.S. 2009. De l'interface à l'approche méthodologique de l'intégration en phonétique et phonologie. *Quel avenir pour les langues et cultures du Gabon*, edited by P. Ondo Mebiame. Libreville: Editions CUI-Gabon.234-247.
- Ndinga-Koumba-Binza, H.S. 2011. Interaction of variables in the Civili vowel duration. *Proceedings of the 17th International Congress of Phonetic Sciences 17-21 August 2011 Hong Kong*, edited by Wai-Sum Lee & Eric Zee. Hong Kong: City University of Hong Kong.1458-1462.
- Ndinga-Koumba-Binza, H.S. 2012. *A phonetic and phonological account of the Civili vowel duration*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Ndinga-Koumba-Binza, H.S. and J.C. Roux 2009a. Perceived duration in length-based Civili minimal pairs. *South Africa Journal of African Languages* 29:216-226.
- Ndinga-Koumba-Binza, H.S. and J.C. Roux 2009b. The representation of vowel duration in Civili dictionaries. *Lexikos* 19:197-206.

- Odden, D. & M. Odden. Kihehe syllable structure. *The syllable: Views and facts*, edited by H. van der Hulst & N.A. Ritter. Berlin/New York: Mouton de Gruyter.417-445.
- Ohala, J.J. 1990. There is no interface between phonology and phonetics: A personal view. *Journal of Phonetics* 22:153-171.
- Ohala, J.J. 1997. The relation between phonetics and phonology. *The handbook of phonetic sciences*, edited by W.J. Hardcastle & J. Laver. Oxford/Cambridge: Blackwell Publishers.674-694.
- Robinson, A.J. 1998. Speech analysis. <http://svr-www.eng.cam.ac.uk/~ajr/SA95> Accessed: 20-02-2007.
- Roux, J.C. 1995. On the perception and production of tone in Xhosa. *South African Journal of African Languages* 15(4):196-204.
- Roux, J.C. & H.S. Ndinga-Koumba-Binza. 2011. Perceived vowel duration in Civili: Minimal pairs and the effect of post-vocalic voicing. *Proceedings of the 17th International Congress of Phonetic Sciences 17-21 August 2011 Hong Kong*, edited by Wai-Sum Lee & Eric Zee. Hong Kong: City University of Hong Kong.1726-1729.
- Ru, P., T. Chi & S. Shamma. 2003. The synergy between speech production and perception. *The Journal of the Acoustical Society of America* 113:498-515.
- Tams A. 1999. Experiments in spoken language.  
<http://www.essex.ac.uk/speech/teaching>. Accessed: 15-08-2001.

---

## **PART 2:**

### LANGUAGE DESCRIPTION AND RESOURCES

---

## CHAPTER 7

# THE ORATURE-GRAMMAR INTERFACE: ON “RHYMES” IN AFRICAN ORAL VERBAL ART

H. Ekkehard Wolff

*Department of African Linguistics, Leipzig University, Leipzig, Germany*  
wolff@uni-leipzig.de

### 1. INTRODUCTION

The purpose of this chapter is to identify and illustrate examples of some of the excessively rich patterns and diverse strategies in African languages which, in genres of traditional verbal art, are used to create “rhymes” as aesthetic highlights within the linear ordering of utterances. By “rhyming strategies” I refer to the wilful usage, by speakers of African languages, of all kinds of repetitive patterns that would qualify as aesthetically motivated parallelisms, henceforth also referred to as “rhymes”. The analysis and description rests on the assumption that mature speakers will attempt to use such rhymes whenever they see an opportunity to do so in the course of their oral production of culturally relevant texts.

In this sense, I use the term “rhyme” with a much broader meaning as compared to the traditional European understanding in terms of only “word rhyme”, i.e. involving mostly vocalic (*assonance*) and/or consonantal (e.g. *alliteration*) parallelisms. In this chapter I will also and predominantly look at repetitive syntactic and morphosyntactic patterns which could also be referred to as “constructional” parallelisms or rhymes. Such rhymes are here considered to be salient features of traditional orature culture which testify to the mastery of the language on the part of the speaker and which contribute to the genuine verbal artistry that is part and parcel of mature adult rhetoric which African societies tend to hold in high esteem. All in the absence of culturally established professional or semi-professional poets or (praise) singers like the *griots* and *imbongi* in West and South African traditions respectively.

For illustrative purposes as much as for limitations of time and space, I will restrict the chapter and discussion to one [Central] Chadic language, Lamang, that I am particularly familiar with, and which is spoken in north-eastern Nigeria somewhat parallel to the international boundary with northern Cameroon<sup>1</sup>. It is the contention of the author that, not the least based on previous findings on the aesthetic features of Lamang verbal art, parallelisms of all kinds, i.e. phonological, morphological, and syntactic parallelisms, together with what I had once termed the “eutonic



structures” of proverbs and riddles, constitute a salient feature of mature adult rhetoric in *Gwàf Làràṅ* as much as in other African languages.<sup>2</sup>

The interface of language and culture, i.e. the relationship between linguistic expressions as manifestations of culture-specific coding strategies related to worldview and observable sociological and cultural features of Lamang society had already been addressed on previous occasions:

- A first attempt to relate certain formal aspects of linguistic expressions to strategies of encoding two different culture-specific notions of “belonging” – both with regard to material possession and kinship relations – is contained in Wolff (1974). Quite clearly, this had nothing to do with *alienable* versus *inalienable* possession, as one might have thought. In the first ever linguistic account of the language (Lukas 1964) which was then known under the name “Hitkalanci”, the German dean of Chadic linguistics, Johannes Lukas (1901-1980), had confessed to his bewilderment about and lack of understanding of, the semantics behind these two different coding strategies for expressions of “belonging” which, as it turned out, reflected sociologically relevant dimensions within Lamang society. In a nutshell: The two coding strategies reflect the double social nature of a member of a Lamang speaking group as individual on the one hand, and member of his/her unilinear descent group on the other.
- Since proverbs and riddles tend to linguistically encapsulate culture-specific aspects of worldview and traditional folklore, a small monograph on Lamang verbal art, focussing on proverbs and riddles, was published a few years later (Wolff 1980) and was subsequently related to the wider context of studying “tonal rhymes” in African languages (Wolff 1998).
- More recently, two papers were published in honour of two leading scholars in African linguistics, in which other aspects of the language-culture interface were explored: (a) by looking at the encoding of features of topography in Lamang grammatical structure (Wolff 2006a), and (b) concerning some ethnohistorical accounts relating to genealogical continuity and discontinuity of Lamang speaking sociocultural groups (Wolff 2006b).

In this vein, the present paper continues a promising line of research that would also allow the culturally and linguistically non-initiated reader to gain some deeper insight into how the culture of some Lamang speaking groups in north-eastern Nigeria works and how this is reflected in linguistic structure.<sup>3</sup> The illustrative examples used and discussed in this paper are taken from the collection of oral literature (Wolff *et al.* 1994) which also provides translations and comprehensive annotations and may serve as an introduction to Lamang culture via their own *verbatim* accounts of some of the salient cultural features of their society.<sup>4</sup>

## 2. STUDYING THE AESTHETICS OF ORAL VERBAL ART

### 2.1. Theoretical and Methodological Background

The very first study of Lamang oral verbal art as presented some 30 years ago (Wolff 1980) had to be placed, in a double sense, into a poorly equipped setting in terms of theoretical and methodological guidance, in particular with regard to “literary” traditions in Africa outside the scope of poetic influence from Arabo-Islamic models. Firstly, “literary linguistics” as it is referred to, for instance, in Fabb (1997) was just beginning to see the light of day in the 1970s but could not yet claim general acceptance as a subfield of both linguistics and literary science, despite visionary contributions by, for instance, Roman Jakobson (1960) and J. S. Petöfi (1972), to mention two works which I found particularly enlightening at the time, contained in a ground-breaking reader edited by J. Ihwe (1972). Secondly, some salient features of the aesthetics of African verbal art had not yet been properly recognised outside Africa, prior to the emergence of first insightful studies from within Africa, such as by D. P. Kunene (1971), L. A. Boadi (1972) and Dalhatu Muhammad (1978). To put it bluntly: When looking at Africa from an outside perspective, traditional African poetry was considered largely devoid (not to say incapable) of, for instance “rhyming” unless metric and/or rhyming patterns were copied from Arabo-Islamic models (cf. Finnegan 1970:74, Greenberg 1949, 1960).

The theoretical and methodological approach to Lamang verbal art chosen in Wolff (1980) had focused on two targets: (a) to establish hard-core linguistic analysis as a valuable tool for analysing aesthetically motivated “artistic” discourse in languages and cultures other than one’s own, and (b) to establish the study of African “oratures” as a central research area for what is called *Afrikanistik* in German academic tradition, and as it was founded during the colonial period in Berlin (since 1885), Leipzig (since 1895) and Hamburg (since 1909). Thirty years later it is good to realise that both targets have been widely achieved, even though “oratory linguistics” (to coin a new term in analogy to “literary linguistics”) has remained somewhat marginal in national *Afrikanistik* as much as in international *African linguistics* and *African Studies*. For a kind of state-of-the-art account of literary linguistics as applied to different human languages and cultures, very much in the vein of what was envisaged in Wolff (1980), cf. Fabb (1997).

2.2. Phonological and Morphological Rhymes

Research on a relatively small corpus of proverbs and riddles in Lamang had shown that these miniature forms of verbal art, made up of syntactic units (“sentences”) which minimally contain one simple clause but may form complex sentences containing several embedded clauses, could be packed with various manipulations of linguistic structure, in addition to semantically highly artistic lexical selection (Wolff 1980). Not surprisingly, the proverbs qualified for the features “compressed and forceful language”, “terseness of expression”, and “shortness”, which characterise proverbs in African languages in general as had been pointed out in Ruth Finnegan’s seminal work on “Oral Literature in Africa” (1970). In addition to Finnegan’s generalisations and coming somewhat as a surprise to experts, however, were the highly complex rhyming structures, both in terms of segmental and suprasegmental (i.e. pitch-related, tonal) structure. Quite obviously, the creation of parallelism of all sorts is one of the major principles that govern the miniature pieces of Lamang verbal art, i.e. proverbs and riddles.

Such parallelism is witnessed for instance in stunning complexities of combined consonant rhymes (*alliteration*), vocalic rhymes (*assonance*), and *tonal rhymes* which I have elsewhere referred to *in toto* as “phonological parallelism” (Wolff 1980:100). See the following example which also illustrates the general pattern of linear division into two or three “lines” of a “proverbial stanza”. For interlinear glossing of the examples the reader is referred to the appendix.

(1) Phonological rhymes: consonantal, vocalic, and tonal parallelisms

Yáyá Yágh	tà Bókò.	Squirrel has given birth to Hyena.
y y y		consonantal rhyme pattern ( <i>alliteration</i> )
a a a	o o	vocalic rhyme patterns ( <i>assonance</i> )
H H H	L L L	<i>tonal rhyme pattern</i> (“ <i>inversion</i> ”)

**Note:** In tale genealogy, *Yaghe* (Squirrel) and *Boko* (Hyena) are siblings (of the same father). This proverb, therefore, comments on a somewhat “pervert” situation or view of things.

The proverbs and riddles further provide evidence for what may be called “lexical” and/or “morphological” parallelisms, or “morphosyntactic rhymes” which are created by appropriate lexical choice. In the following example, the proverb exploits the formal morphological identity of three occurrences of a noun each with possessive 2<sup>nd</sup> person singular marking, i.e. suffixing the morpheme sequence **-àa-ghà** (with final **à** of the possessive pronoun automatically deleted if it is followed by another vowel).

(2) Morphological rhyme (in combination with phonological rhymes)

Hà m- <b>àa-gh</b> -úw ná	márd- <b>àa-gh</b> -é	m- <b>àa-ghà</b>	(If) you don’t have your (own) mother, it is your stepmother (who will be) your mother.
h gh	gh	gh	<i>alliteration</i>
m	m	m	<i>alliteration</i>
L L		L L	<i>tonal rhyme pattern</i> (“ <i>bracketing</i> ”)

**Note:** Pre-Islamic and Islamic members of the Lamang speech communities allow polygyny. Women in a polygynic household, in principle, share responsibility for all children of the common husband. However, the notion of “bad step-mother” is also familiar to *Gwàd Lámàn* speakers. The proverb makes reference to the need to adjust in life to both good and bad circumstances.

### 3. SYNTACTIC RHYMES

#### 3.1. Syntactic Rhyme in Oral Proverbs and Riddles

As could be assumed and was explicitly anticipated (Wolff 1980:100), instantiations of syntactic parallelism would – most likely – occur in longer pieces of oral verbal art, i.e. prose texts such as stories for entertainment, narrative and descriptive texts etc. However, until this day no study had been undertaken to check on the validity of such assumption for Lamang oral prose texts. The present contribution attempts to make the picture complete by providing and illustrating examples of syntactic parallelism. Since we recognise syntactic parallelism for its artistic quality, it will be henceforth also referred to as “syntactic rhyme”.

We will, however, start by looking at syntactic rhymes that had already occurred in the miniatures of verbal art, i.e. proverbs and riddles which – in addition – often carry further aesthetic features in terms of phonological rhymes, in particular so-called “eutonic” structures. (Note that in the examples following, the rhyming parts of the utterances are “boxed”, the non-rhyming parts of the examples are left outside the “boxing”.)

(3)	<i>Másúlá ñ tòn ndàrdí</i>	<i>tálálóñ ñ tòn éwé.</i>	(He has) weak buttocks, (but has a) big mouth
	H H H L L (L) H	H H H L L (H) H	<i>tonal rhyme</i> pattern ("repetition")

**Note:** The Lamang consider firm and round buttocks as a sign of health, strength, and beauty. The proverb makes reference to a person whose "big mouth", i.e. like inappropriately speaking out in public, is not matched by his/her social (and intellectual) standing. (cf. the corresponding lines in the song *Big hat, no cattle* by one of my personal favourite singers, Randy Newman (1999))

(4)	<i>Síhá lá-hápódf tá gháñ</i>	<i>síhá lá-ghábárùkù.</i>	Some eat the beans, others pass the wind.
	<i>h h gh</i>	<i>h gh</i>	<i>alliteration</i>
	H H H H H H H	H H H H H / L L	<i>eutonic coda contrast</i>

**Note:** This proverb is a comment on the lack of justice in the world where some members of society profit, in an unjustified manner, from the efforts of others.

(5)	<i>ñ zó</i>	<i>ñ mònò.</i>	(You are) eating as (you are) doing.
	<i>ñ verb-o</i>	<i>ñ verb-o</i>	morphemic rhyme pattern (preposition <i>ñ</i> + verb with nominaliser -o)

**Note:** The proverb refers to the idea that "you eat as much as you work", i.e. that you yourself are responsible for your own doings and its consequences in good or bad.

(6)	<i>Dzòv t ínàa zà bé bàlà-w</i>	Finding something to eat that's not difficult,
	<i>dzòv tà bàgáa záatè bàlà</i> <i>ká figw.</i>	finding a place to eat is difficult, says the young chick.

**Note:** The proverb refers to the experience that one has to compete - if not fight - for one's own position in society.

The following example is a riddle that calls for the identification of two related concepts; the riddle comes in the shape of a syntactic rhyme and the expected answer also displays rhyming in terms of morphosyntactic structure. Note that and how the riddle is introduced by a ritualised two-line dialogue between the person that asks the riddle and the person offering to answer it.

(7) *àrgwàanzí.*

I am a riddle.

*Súutì!*

Come and let me taste it!

<i>Yághw, dàd t éwáa wáatàghà.</i>	Salute to you, elder at the front entrance.
<i>Yáwá, gwàsa t úhóláa hgà.</i>	Salute to you too, son-in-law behind the house.

ANSWER:

<i>ḍáfá-yá.</i>	It is food.
<i>ghúvâ-yà.</i>	It is the excrements.

**Note:** The thatched front entrance hut of a compound (*wáatàghà* lit. "mouth-of residence") is the "parlour" where the owner of the compound receives other elders on formal or informal occasions. Therefore, it is the favourite place for elders to sit. An elder is greeted by a younger person as *dàdà* "father". Like in many cultures, the latrine is located "behind the house" or compound and is often referred to, euphemistically, by indicating this location (cf. also the Hausa expression *bayan gida* "toilet; lit. back-of compound"). This riddle, besides being considered extremely funny, like most riddles, is usually put to children and serves educational purposes: It refers to the highly important affinal (in contrast to consanguine) kinship relations as well as to the nourishment cycle, i.e. the causal relationship between eating and defecating.

### 3.2. Syntactic Rhyme in Oral Narrative and Descriptive Prose

In the following section, I will present examples and illustrations from two different genres of Lamang prose orature. The first piece of oral literature to be considered represents a so-called dilemma story. Dilemma stories may be quite short and are told, not the least, for amusement but, almost in passing they tend to refer to moral values and the difficulty, if not impossibility, of judging one moral value in society higher than another. The narrator typically develops a story at the end of which the protagonist is caught in a terrible dilemma where all available options appear to be equally unacceptable because they would violate salient

cultural norms. At the end of such a dilemma tale, the listener is asked, implicitly or explicitly, what (s)he would do if (s)he were in the place of the protagonist. The dilemma tale that we will look at here is the one of "The squirrel-hunter and his children" (Wolff *et al.* 1994:228ff., 348f). In this story, a certain boy is ill-treated by his own father, finds a benefactor who offers him the best of all possible ways of living, and is then forced to take a fatal decision against either his own father, or his benefactor, or himself. Since neither the dilemma scenario as such nor the story as a whole are of concern here, I will restrict the discussion to the occurrence of what I consider syntactic rhymes in this specimen of a particular genre of Lamang oral tradition. The first occurrence of syntactic rhyming follows the introduction to the story by describing how the father and his sons hunt a ground-squirrel (*yaghe*). They detect a squirrel and start to chase it.

- (8) *má st dèghúlhá,*  
the boys closely behind,

<i>lāa yāgh dā mā-nā      pātsā,</i>	<i>ndáhaŋ;</i>
<i>lāa yāgh dā mā-yā,</i>	<i>ndáhaŋ.</i>
the squirrel ran into some fresh green here,	they followed;
the squirrel ran into that,	they followed.

As the story develops, the ill-tempered father of the boys almost beats one of his sons to death because the boy had failed to catch that squirrel. The boy was found half-dead by a rich trader who took him home and made him his partner in the trade of hides and skins. One day the boy's father entered the market where his son was doing very good business with trading hides and skins. He saw him sitting there as a well-established trader, shouted at him and tried to force him to come home with him to be a poor squirrel hunter again like his brothers, saying,

- |     |  |   |
|-----|--|---|
| (9) | <i>úzíná màghà</i>                     | your brothers (in the female line)      |
|     | <i>nd' úzíná dàghà,</i>                | and your brothers (in the male line),   |
|     | <i>yághě tkóshàn;</i>                  | they hunt squirrel(s);                  |
|     | <i>skwóghà mtáká</i>                   | coming home (from) bush                 |
|     | <i>hàhàd' ùndà tks t' yágh hwatáfá</i> | there's some who catch five squirrel(s) |
|     | <i>hàhàd' ùndà tks t' yágh ùfádá</i>   | there's some who catch four squirrel(s) |
|     | <i>máhàn ná, ...</i>                   | [PRS.3PL], ...                          |

The boy's benefactor intervenes in the arising fight between father and son, offers the boy his sword and puts three options to him to solve the dilemma that the boy is now in – and the story ends with a beautiful triple syntactic rhyme sequence over three lines (which for reasons of representation shall be indicated as columns A-B-C under (10); each of the lines 1., 2., 3. should be read in the sequence A-B-C):

(10) àghǎŋ kátsákárǎná, àaspá!

Here (is) this sword, take it!

A		
1.	dátsá- ghǎŋ	-ka t' iyo;
2.	á'á, dátsá- ghǎŋ	-kà tò dàghà;
3.	á'á, dátsá- tá	-ká tǎ ghǎŋ tò ghǎŋàghà;

B		
1.	tsáká ghǎŋ	t' íyó
2.	tsáká tò ghǎŋ	tò dàghà
3.	tsáká ghǎŋ-àghá	

C		
1.	gúká dzághà	ndà dàghà;
2.	gúmǎŋ dzághà	ndà kàghà;
3.	gúy dzághà,	
	gú dàgh	dzághà

Line 1 A-B-C translates as:

"You could decapitate me; you decapitate me, and you go home with your father."

Line 2 A-B-C translates as:

"If not, you could decapitate your father; you decapitate your father, and you and I go home."

Line 3 A-B-C translates as:



"If not, you could decapitate yourself; you decapitate yourself, and I go home, and your father goes home."

The second example from recorded prose discourse belongs to an entirely different text genre; as a matter of fact, the text in itself constitutes an innovative genre in Lamang orature. It is a "juridical" text on inheritance rules among a group of Lamang speakers that was first written down, in a non-standardised ad hoc orthography, by a then student of Law at the University of Zaria who, for recording purposes, read it on tape from his notes. The text is the first ever recorded "academic" text in Lamang oratory history. This text is much longer (in comparison with, for instance, a dilemma story). This prose text on Lamang inheritance rules was chosen because it stems from a different male speaker with a different educational background. It also shows that the use of syntactic rhymes is not an idiosyncratic and somewhat whimsical feature of just one particular speaker or narrator, i.e. for instance that of the narrator of the dilemma story. And, of course, as was pointed out, it constitutes a prose text totally outside the received orature traditions.<sup>5</sup>

Given the rather technical "legalistic" content of the prose text on Lamang inheritance rules (*wunaa laman* [Wolff *et al.* 1994:134-141, 291-296]), it is not surprising to find that most occurrences of syntactic parallelism would also qualify as "listings". We note, however, that such listings make use of full and parallel syntactic patterns which correspond to the syntactic rhymes in the entertaining dilemma story of a different genre as illustrated further above.

(11) *vita hà ínàa dìn kádà gùlèṅ-úwó*

If, however, the thing(s) of his father are not plentiful

<i>kúkálá Yágà,</i>	<i>Yágà takes (his share)</i>
<i>kúkálá Slàgháyá,</i>	<i>Slàgháyá takes (his share)</i>
<i>kúkálá dábà Yágà,</i>	<i>dábà Yágà takes (his share)</i>
<i>kúkálá dábà Slàgháyá.</i>	<i>dábà Slàgháyá takes (his share)</i>

**Note:** *Yágà*, *dábà Yágà* ("the one following *Yágà*"), *Slàgháyá* (or: *Makadzi*), *dábà Slàgháyá* ("the one following *Slàgháyá* or: *Makadzi*") are fixed terms in Lamang culture which, following birth order and marital status of their respective mothers, designate certain children of a deceased male person in terms of inheritance rights.

(12) *vita gùlèṅ māràkwá t slálá gùlò tálá úzàṅ médé, úzàṅ zùgùné gùlèṅ,*

if, again, (there is) a first wife with one child, a male child, that is,

*hà úndá sɪd mɛd -úwó* there is nobody else of hers

*hà mákwìn -ùwò* there is no daughter of hers

*hà úzàṅá sɪd -uwo...* there is no (other) child ...

**Note:** *márákwá t slálá gùlò* 'wife on the big kitchen' is a descriptive term for the "first wife" of a husband.

- (13) *vita mámtá ùndù* if a person dies  
*zlàváb t úzàṅá zùgún -úwó* (and if) he does not leave a son  
*zlàváb t úzàṅáa mín -uwo* (and if) he does not leave a maternal brother  
*zlàváb t úzínáa dín -ùwò* (and if) he does not leave paternal brothers  
*né gòlò dàmònò?* What will be done?  
*Warka íná zlàvàḍe ná* The only thing he leaves (being)  
*zlàvàzlá t dáslǎṅìni* he leaves a paternal uncle of his  
*zlàvàzlá t míhàmbìni* he leaves his wives  
*ndà gùlég* and also  
*zlàvàzlá t ùzàṅáa dáslǎṅìni...* he leaves a child by his paternal uncle...

(14)	<i>vita hà úzínáa mín-úwó ná</i>	If there are no maternal brothers
	<i>vita hà gùlég úzínáa dín-úw ná</i>	if there also are no paternal brothers

<i>úzínáa dīne thálo tà míhàmbìni</i>	the paternal brothers inherit his wives
<i>úzínáa dáslǎṅhé dáhál tà-hǎṅ</i>	the paternal cousins will inherit them

(15)	<i>Yágàná: úvàh-há hkán-é dǎvlǎŋtǎlò</i>	As for Yágà: he will be given three farms
	<i>ḍábà Yágà: úvàh-há hés-é dǎvlǎŋtǎlò</i>	ḍábà Yágà: he will be given two farms
	<i>Slàgháyá: úvàh-há úfád-é dǎvlǎŋtǎlò</i>	Slàgháyá: he will be given four farms
	<i>ḍábà Mákàdzí: úvàh-há hkán-é dǎvlǎŋtǎlò</i>	ḍábà Mákàdzí: he will be given three farms

(16)	<i>Yágà: úvàh-há hkáná ŋ rét-é dǎvlǎŋtǎlò</i>	Yágà: three and a half farm one will give (him)
	<i>ḍábà Yágà: úvàh-há hkán-é dǎvlǎŋtǎlò</i>	ḍábà Yágà: three farms one will give (him)
	<i>Slàgháyá: úvàh-há hésá ŋ rét-é dǎvlǎŋtǎlò</i>	Slàgháyá: two and a half farm one will give (him)
	<i>ḍábà Mákàdzí: úvàh-há hkán-é dǎvlǎŋtǎlò</i>	ḍábà Mákàdzí: three farms one will give (him)

#### 4. CONCLUSION AND LOOKING AHEAD

The syntactic analysis of recorded African oral prose texts, as illustrated with examples from Lamang orature in North-eastern Nigeria, may testify to frequent occurrence of “constructional” parallelism of sorts, largely independent of text genre and speaker. Such kinds of “syntactic rhyme” had already been observed in miniature genres of verbal art, i.e. proverbs and riddles, in the same language. Quite obviously, proverbs and riddles as much as prose orature use the same stylistic strategies in order to create artistic verbal expression and, thereby, testify to the rhetoric competencies and maturity of the individual speaker.

Since the corpus used constitutes the first ever recorded and transcribed prose texts in this language (the recordings date from extended periods of linguistic field work in 1968/69 and 1973/74), one may safely assume that they reflect rhetoric strategies

that characterise Lamang traditional orature as a whole. It is the author’s contention that, if studied thoroughly, similar instances and comparable strategies of aesthetic parallelism, or rhymes, will be found to pertain in many African languages, and possibly even in languages elsewhere around the world. Therefore, the present paper set out to address the orature-grammar interface as alluded to in the title not only for the particular language that was chosen for illustration, i.e. Central Chadic Lamang, but also in a much more general sense.

For the language used for illustration, it remains to be seen whether and how more recent writings in Lamang deviate to any considerable extent from the model provided by the two principal narrators, the late Alhaji Abdullahi Ndaghra and Chief Justice Ibrahim Garndawa, who both belong to a generation that was linguistically and culturally socialised before and during the time of World War II, i.e. some 70-60 years ago. Their rhetoric competence is well documented in the two books *Ina Lamang* (Wolff & Ndaghra 1992) and *Our People’s Own* (Wolff *et al.* 1994), which were the first books ever to be written and published in Lamang. The author is not aware of any further prose texts locally produced as a consequence of the launching ceremony of *Ina Lamang* at the Emir of Gwoza’s palace in 1992.<sup>6</sup>

Sporadic and anecdotal evidence available to me indicates that among the younger generations of Lamang speakers, passive competence still largely allows them to understand the *Ina Lamang* prose from the 1960s and 1970s, but that their active command of the intricacies of the language is rapidly fading. Clearly, Lamang has become an endangered language that is no longer fully transmitted to following generations. Younger members of the originally Lamang speaking community are increasingly shifting to Hausa in the north, and Fulfulde in the south of the language territory. Hausa in particular is becoming the widest spread lingua franca in most parts of Northern Nigeria and tends to become the dominant and preferred language of the youth, not the least because of the effects of formal education and work migration into the mainly Hausa speaking urban centres of modern Northern Nigeria.

On a somewhat sad note and speaking from the vantage point of richness and diversity of human creativity and expression, we therefore need to accept the fact that the increase of mobility and Western type education almost inevitably leads to the impoverishment if not loss of both cultural and linguistic achievements of humankind – unless more efforts and means are invested into the empowerment of African and other endangered languages worldwide, not the least in order to slow down, if not neutralise, the sweeping effects of linguistic globalisation. But that would be a topic for another paper!

## ENDNOTES

1. For convenience sake the term Lamang is used to refer indiscriminately to the language as well as to the various groups of speakers of *Gwàḍ Làmàṅ* (lit. "Language [of] Our People") even though they do not constitute an "ethnic" group of sorts: They do not share the idea of a common descent despite speaking the same language. The speakers of *Gwàḍ Làmàṅ* rather represent a number of highly localised groups each of which recognises, first of all, unilinear descent. They trace their origins to independent migrational histories of different patrilinear descent groups which form the basis of their social organisation (cf. Wolff *et al.* 1994:1-29; Wolff 2006b). Speakers of this language may number between 40 - 50 000. *Gwàḍ Làmàṅ* must be considered an "endangered language" due to ongoing language shift to the dominant *linguae francae* in the area, i.e. *Hausa* and *Fulfulde*, on the part of many members of the younger generation.
2. I could cite largely anecdotal evidence from discussions on "undiscovered" rhyming features in "oratures" in various parts of Africa: As a rule and after having given a talk on the aesthetics and "euphonic structures" in some African languages that I am familiar with, speakers of several other African languages in various audiences across the continent would approach me in order to confirm the existence of similar features also in their own traditions of oral verbal artistry which they had not realised to exist prior to hearing my talk. Quite unfortunately, the linguistics of verbal art is hardly ever studied in any detail in language documentation and description. Likewise, serious "literary" research into the linguistics of oratures in African languages is far too rarely attempted, both for minority and even majority languages. It is a few notable exceptions that would confirm this diagnosis as the rule.
3. For details of linguistic structure of *Gwàḍ Làmàṅ*, the reader is referred to the author's monographic description (Wolff 1983).
4. The original *Gwàḍ Làmàṅ* texts, without translations and annotations, were published for non-commercial usage and distribution in the Lamang speaking area as a first contribution to literacy and post-literacy in this language (Wolff and Ndaghra 1992). The support of the German Ministry for Foreign Affairs for this project is once more gratefully acknowledged.
5. Note that for each text, narrators are identified, as is the date of recording, in the published compilations of texts (Wolff & Ndaghra 1992, Wolff *et al.* 1994). While the dilemma story above was told by Alhaji Abdullahi Ndaghra [1928-1978], the "juridical" prose text was compiled by the later Chief Justice Ibrahim Garndawa [born 1938].
6. It was on this occasion of the launching of *Ina Lamaṅ* that the author was "turbaned" and bestowed the title of *Midalaa Lamaṅ* ("Commander of Our People") by the Emir of Gwoza, and another traditional title was bestowed, posthumously, on the late Alhaji Abdullahi Ndaghra. This very rare ceremony (the Emirate had

only once before ever bestowed a traditional title on a member of its own community) was formally declared to serve as official recognition of our joint efforts to document the languages and cultures of the Gwoza Emirate, and thus rescue them from oblivion.

## REFERENCES

- Boadi, L.A. 1972. The language of the proverb in Akan. *African Folklore*, edited by R. H. Dorson. Bloomington: Indiana University Press.183-191.
- Dalhatu, M. 1978. The two facets of rhyme in Hausa poetry: syllabic and tonal. *Harshe* 1:6-18.
- Fabb, N. 1997. *Linguistics and Literature*. Oxford: Blackwell Textbooks in Linguistics.
- Finnegan, R. 1970. *Oral Literature in Africa*. Oxford: The Oxford Library of African Literature.
- Greenberg, J.H. 1949. Hausa verse prosody. *Journal of the American Oriental Society* 69:125-135.
- Greenberg, J.H. 1960. A survey of African prosodic systems. *Culture in history: Essays in honour of Paul Radin*, edited by S. Diamond. New York: Columbia University Press.925-950.
- Ihwe, J. (Ed.) 1972. *Literaturwissenschaft und Linguistik*. Vol.1. Frankfurt a. M.: Fischer Athenäum Taschenbuch Verlag.
- Jakobson, R. 1960. Linguistik und Poetik. *Literaturwissenschaft und Linguistik*, edited by J. Ihwe. Vol.1. Frankfurt a. M.: Fischer Athenäum Taschenbuch Verlag.99-135.
- Kunene, D.P. 1971. *Heroic Poetry of the Basotho*. Oxford: Clarendon Press.
- Lukas, J. 1964. Das Hitkalanci, eine Sprache um Gwoza (Nordostnigerien). *Afrika und Übersee* 48:81-114.
- Newman, R. 1999. *Bad Love*. CD published by Randy Newman Music (ASCAP).
- Petőfi, J.S. 1972. Zur strukturellen Analyse sprachlicher Kunstwerke. *Literaturwissenschaft und Linguistik*, edited by J. Ihwe. Vol.1. Frankfurt a. M.: Fischer Athenäum Taschenbuch Verlag.229-244.
- Wolff, H.E. 1974. Sprachliche Manifestationen des Kollektiv-Denkens der Laamang (Nordostnigeria). *Zeitschrift der Deutschen Morgenländischen Gesellschaft* Supplement II:660-671.
- Wolff, H.E. 1980. *Sprachkunst der Lamang. Stil, Bedeutung und poetische Dimension in zwei Genres oral tradierter Ein-Satz-Literatur*. Glückstadt: J. J. Augustin.

- Wolff, H.E. 1983. *A Grammar of the Lamang Language (Gwàd Lámàŋ)*. Glückstadt: J. J. Augustin.
- Wolff, H.E. 1998. *Afrikanische Sprachminiaturen. Zur formalen Ästhetik von Kleinformen afrikanischer Sprachkunst unter besonderer Berücksichtigung ihrer Tonalität*. ULPA – University of Leipzig Papers on Africa, Languages and Literatures 5. Leipzig: Institut für Afrikanistik.
- Wolff, H.E. 2006a. Encoding topography and direction in the verbal system of Lamang and Hdi (Central Chadic). *West African Linguistics: Papers in Honor of Russell G. Schuh*, edited by Paul Newman and Larry M. Hyman. Studies in African Linguistics, Supplement 11. Columbus, Ohio: Dept. of Linguistics and the Center of African Studies, Ohio State University. 221-250.
- Wolff, H.E. 2006b. Genealogical discontinuity and recontinuity in Hidkala oral traditions. *Africa in the Long Run (Festschrift in the Honour of Professor Arvi Hurskainen)*, edited by L. Harjula and M. Ylänkö. Studia Orientalia 103. Helsinki: The Finnish Oriental Society. 111-129.
- Wolff, H.E. & A.A. Ndaghra (postum) and Eleonore Adwiraah. 1994. *Our People's Own (Ina Lamaŋ). Traditions and Specimens of Oral Literature from Gwàd Lamaŋ Speaking Peoples in the Southern Lake Chad Basin in Central Africa*. Hamburg: Research and Progress.
- Wolff, H.E. & A.A. Ndaghra (postum). 1992. *Ina Lamaŋ. Documents of Oral Traditions in Gwàd Lamaŋ. Collected in the Gwoza Area of Borno State, Nigeria*. Hamburg: Research and Progress.

## APPENDICES

### Appendix I: Conventions and Abbreviations Used in Interlinear Glossing

The symbols and abbreviations for grammatical categories used in the interlinear glossing are listed below. The glossing conventions used are inspired by the “Leipzig Glossing Rules” as suggested by the Dept. of Linguistics of the Max Planck Institute for Evolutionary Anthropology, Leipzig. Since not all of the terms were used in *A Grammar of the Lamang Language (Gwàd Lámàŋ)*, page numbers are given in parenthesis to refer the reader to the relevant sections of the grammatical description in Wolff (1983).

- |         |   |
|---------|---|
| [ ]     | categories with no direct morphological marker (e.g. PRF)   |
| 1, 2, 3 | 1 <sup>st</sup> , 2 <sup>nd</sup> , 3 <sup>rd</sup> person (subject function when attached to verb) |

AUTOBEN	autobenefactive (pp. 119f)
COLL	collective association (pp. 194f)
COP.DEF	copula with definite NP (pp. 88, 218ff)
CPL.NOM	completive verbal nominaliser (pp. 124f, 131f)
CS	clause-final subordination marker (pp. 190, 239-250, 258)
DEF	definite marker (pp. 87f)
DET.NEAR	near demonstrative (pp. 203ff)
DET.FAR	distant demonstrative (pp. 203ff)
EXT	verbal derivative extension (pp. 109-125)
FOC	term focus marker (pp. 256ff)
FUT	future (pp. 179f)
GEN	genitive marker (pp. 95ff, 191ff)
IMP	imperative (pp. 174-177)
INCL	inclusive category of personal pronouns (pp. 82-84)
ITER	iterative-durative (p. 170)
NARR	narrative (pp. 168f)
NEG	negation marker (pp. 172ff, 251ff)
NOM	verbal nominaliser (pp. 125-130)
OBJ	direct object marker (pp. 229ff)
PLACT	pluractional verb base extension (pp. 107-109)
PREP	preposition (pp. 214ff)
PRES	presentative expression (“here you are”, “here is”; pp. 222f)
PRF	perfect(ive) aspect (pp. 133ff, 154ff, 160ff)
PRS	disjunctive personal pronoun (pp. 84f)
PL	plural
POSS	possessive personal pronouns (pp. 85, 95ff, 191ff)
QUOT	quotative (pp. 235-238)
REFL	reflexive (as verbal extension: pp. 121f)
SG	singular



SUBJ	subjunctive (pp. 140ff)
STATE	verbal noun of state marker (p. 130)
TOP	topicalisation marker (pp. 188f, 258)
UHR	unspecific human referent (pp. 90-94, 199ff)

## Appendix II: Examples with Interlinear Glossing

In the following transcriptions of the examples which are used in the body of the chapter, the frequently occurring syncopation and apocopation of vowels in spoken discourse is “undone”, i.e. the syncopated or apocopated vowels are restored (without their only partly predictable tones) in parentheses in order to allow for a better identification of the “full form” of the respective lexical item or morpheme as far as this full form is known from elicitation or other text materials. This does not, however, apply to vowels which are deleted by rule when followed by another vowel. Note that phonetic centralisation of phonemic /a/ to [ə] remains transcribed as such, as is the insertion of pro- and epenthetic vowels in the environment of phonemic consonant clusters.

Note that a “Dictionary of the Lamang Language” is still in the making, its pending completion hinges exactly on the problems sketched out above of identifying “full” forms that could be best used as lexical entries.

- (1) *yá-yá* *yágh(e)* *tà* *Bòkò*  
give.birth-give.birth[PRF.3SG] groundsquirrel OBJ Boko  
Squirrel has given birth to Hyena.
- (2) *hà* *m-àa-gh-úw(o)* *ná*  
exist[3SG]/NEG mother-GEN-2SG.POSS-NEG CS  
*mórd-àa-gh-é* *m-àa-ghà.*  
stepmother-GEN-2SG.POSS-FOC mother-GEN-2SG.POSS  
(If) you don't have your (own) mother, (it is) your stepmother (= father's  
cowife) (who will be) your mother.
- (3) *másúlá ñ* *tàv(e) ndàrdí* *táláláñ ñ* *tàv(e) éwé*  
weak PREP way buttocks strong PREP way mouth  
(He is) weak by buttocks, (but has a) big [by] mouth.
- (4) *síhá* *lá-hápód(-o)* *tá gháñ*  
different UHR.PL-chew[NOM] OBJ bean(s)

*síhá lá-ghábár-ùkù.*  
different UHR.PL-flatulent.belly-STATE

Some eat the beans, some others pass the wind.

- (5) *ɲ z-ó ɲ mòn-ò*  
PREP eat-NOM PREP do- NOM

(You are) eating as (you are) doing.

- (6) *dzòv(-ò) t(á) ín-áa z-ò b-é bàlà-w(o)*  
get- NOM OBJ thing- GEN eat- NOM NEG- FOC difficult[3SG]-NEG  
*dzòv(-ò) tà bàg-áa z-áa`-t-e bàlà*  
get-NOM OBJ place-GEN eat-EXT-CPL.NOM-FOC difficult[3SG]  
*ká figw(i).*  
say[3SG] chick

Finding something to eat (is) not difficult, finding a place to eat (is) difficult, says the young chick.

- (7) *yághw(è), dàd t(á) éw-áa wáatàghà*  
salute, father PREP mouth-GEN entrance.of.compound  
*Yáwá, gwàsa t(á) úhál-áa hga*  
salute, son-in-law PREP back- GEN compound

Salute to you, elder at the front entrance.  
Salute to you too, son-in-law behind the house (at the latrine).

*dáf-áyá ghúv-àya.*  
mush- COP.DEF excrements- COP.DEF

That is food – that is excrements.

- (8) *má st(o) dàghúl-há,*  
PREP bottom young.man-PL  
*l-áa yágh dá má-ná pátsá ndá-háj*  
go- EXT squirrel PREP PREP-DET.NEAR fresh.green follow-3PL  
*l-áa yágh dá má-yà ndá-háj.*  
go- EXT squirrel PREP PREP-DET.FAR follow-3PL

The boys closely behind, the squirrel ran into some fresh green here, they followed; the squirrel ran into that, they followed.

- (9) *úzín-á(a) m-à(a)-ghà*  
children-GEN mother-GEN-2SG.POSS

*nd(a) úzín-á(a) d-à(a)-ghà,*  
and children-GEN father-GEN-2SG.POSS

*yágh-é t(a)-kós-háŋ*  
squirrel- FOC ITER-catch.PLACT[NOM]-3PL

*skwó-ghà mtáká*  
come- EXT.home bush

*hàhà-d'(e) únd-à t(a)-ks(-o) t(à) yágh(e) hwtáfá*  
exist-3SG person-DEF ITER-catch(-NOM) OBJ squirrel five

*hàhà-d'(e) únd-à t(a)-ks(-o) t(à) yágh(e) ùfádǎ*  
exist-3SG person-DEF ITER-catch(-NOM) OBJ squirrel four

*má-háŋ ná, ...*  
PREP-3PL CS

Your brothers (in the female line) and your brothers (in the male line), they hunt squirrel(s). (When) coming home (from) bush, there's some who catch five squirrel(s), there's some who catch four squirrel(s).

- (10) *àgháŋ kátsákàr-à-ná àa-spá*  
PRES sword-DET.NEAR [SUBJ/IMP-]EXT-hold

Here (is) this sword, take it!

*dá-tsá-ghaŋ-ka t(à) iyo*  
FUT-cut-head -2SG OBJ PRS.1SG

You could decapitate me;

*tsá-ká ghàŋ t(à) íyó*  
cut-2SG head OBJ PRS.1SG

you decapitate me

*gú-ká dzá-ghà ndà d-à(a)-ghà*  
NARR-2SG go-EXT.home with father-GEN-2SG.POSS

and then you go home with your father;

*á'á dá-tsá-ghaŋ-kà t-à d-à(a)-ghà;*  
no FUT-cut-head-2SG OBJ father-GEN-2SG.POSS

if not, you could decapitate your father;

*tsá-ká t-à ghàŋ t-à d-à(a)-ghà*  
cut-2SG OBJ head OBJ father-GEN-2SG.POSS

you decapitate your father

*gú-máŋ dzá-ghà ndà kàghà*  
 NARR-1PL.INCL go- EXT.home with PRS.2SG

and you and I go home;

*á'á dá-tsá-tá-ká tá ghàŋ tà ghàŋ-à(a)-ghà*  
 no FUT-cut-CPL.NOM-2SG OBJ head OBJ head-GEN-2SG.POSS(REFL)

if not, you could decapitate yourself;

*tsá-ká gháŋ-à(a)-ghá*  
 cut-2SG head-GEN-2SG.POSS

you decapitate yourself

*gú-y(i) dzá-ghà*  
 NARR-1SG go-EXT.home

*gú d-à(a)-gh(à) dzá-ghà*  
 NARR[3SG] father-GEN-2SG.POSS go- EXT.home

and I go home and your father goes home

- (11) *vita hà íná-a d-in(i) kádà gùlér-wó*  
 if exist[3SG] thing-GEN father-3SG.POSS plenty also-NEG

If, however, the thing(s) of his father are not plentiful

*k-ú-kálá Yágà*  
 take-AUTOBEN-take[PRF.3SG] Yágà,

Yágà takes (his share)

*k-ú-kálá dábà Yágà*  
 take-AUTOBEN-take[PRF.3SG] dábà Yágà,

dábà Yágà takes (his share)

*k-ú-kálá Slàgháyá*  
 take-AUTOBEN take[PRF.3SG] Slàgháyá

Slàgháyá takes (his share)

*k-ú-kálá dábà Slàgháyá*  
 take-AUTOBEN-take[PRF.3SG] dábà Slàgháyá

dábà Slàgháyá takes (his share)

- (12) *vita gùlér(e) márakwá t(a) slálá gùlò tálá úzàŋ(a) mé-dě*  
 If also wife PREP kitchen big one child for-her

if, again, (there is) a first wife with one child

*úzàŋ-á zùgùn-é gùlég(e)*

child- DEF male- FOC also

a male child in particular

*hà únd-á sìd(i) mé-d-úwó*

exist[3SG] person- DEF certain for-her-NEG

there is nobody else of hers

*hà mákw-ìn-ùwò*

exist girl-3SG.POSS-NEG

there is no daughter of hers

*hà úzàŋ-á síd-uwo*

exist child- DEF certain-NEG

there is no(other) child

- (13) *vita má-mtá ùndù*  
if die-die[PRF.3SG] person

if a person dies

*zlà-và-b t(à) úzàŋ-á zùgùn-úwó*

leave-REFL-EXT[3SG] OBJ child- DEF male-NEG

(and if) he does not leave a son

*zlà-và-b t(à) úzàŋ-áa m-ìn-uwo*

leave-REFL-EXT [3SG] OBJ child- GEN mother-3SG.POSS-NEG

(and if) he does not leave a maternal brother

*zlà-và-b t(à) úzín-áa d-ín-ùwò*

leave-REFL-EXT[3SG] OBJ children- GEN father-3SG.POSS-NEG

(and if) he does not leave paternal brothers,

*n-é gò-lò dá-mòn-ò*

what-FOC QUOT-3.INCL FUT-do-NOM

what will be done?

*Warka íná zlà-và-dé ná*

only thing leave-REFL-3SG CS

The only thing he leaves (being)

*zlà-và-zlá* *t(à)* *dásláŋ-ini*  
 leave-REFL-leave[PRF.3SG] OBJ paternal.uncle-3SG.POSS

he leaves a paternal uncle of his

*zlà-và-zlá* *t(à)* *míhá-mb-ini*  
 leave-REFL-leave[PRF.3SG] OBJ wives-COLL-3SG.POSS

he leaves his wives

*ndà gùléŋ(e)*

and also

and also

*zlà-và-zlá* *t(à)* *ùzàŋ-áa* *dásláŋ-ini...*  
 leave-REFL-leave[PRF.3SG] OBJ child- GEN paternal.uncle-3SG.POSS

he leaves a child by his paternal uncle...

- (14) *vita hà* *úzín-áa* *m-in-úwó* *ná*  
 if exist[3SG] children-GEN mother-3POSS-NEG CS

If there are no maternal brothers

*úzín-áa* *d-in-e* *t(a)-háI-ó* *tà* *míhá-mb-ini*  
 children-GEN father-3POSS-FOC ITER-take-NOM OBJ wives-COLL-3POSS

the paternal brothers inherit his wives

*vita hà* *gùléŋ(e)* *úzín-áa* *d-in-úw* *ná*  
 if exist[3SG] also children- GEN father-3POSS-NEG CS

if there are no paternal brothers

*úzín-áa* *dásláŋ-h-é* *dá-háI(-o)* *tà-háŋ*  
 children-GEN paternal.uncle-PL-FOC FUT-take(-NOM) OBJ-3PL

the paternal cousins will inherit them.

- (15) *Yágà-ná* *úväh-há* *hkán-é* *dá-vIá-ŋ-tà-lò*  
*Yágà*-TOP farm-PL three-FOC FUT-give-EXT-CPL.NOM-3INCL

As for Yágà: he will be given three farms,

*dábà Yágà* *úväh-há* *hés-é* *dá-vIá-ŋ-tà-lò*  
*dábà* Yágà farm-PL two-FOC FUT-give-EXT-CPL.NOM-3INCL

*dábà* Yágà: he will be given two farms,

*Slàgháyá úvâh-há úfâf-é dá-vlâ-ŋ-tâ-lò*

Slàgháyá farm-PL four-FOC FUT-give-EXT-CPL.NOM-3INCL

Slàgháyá: he will be given four farms,

*ḍábà Mákàdzí úvâh-há hkân-é dá-vlâ-ŋ-tâ-lò*

ḍábà Mákàdzí farm-PL three-FOC FUT-give-EXT-CPL.NOM-3INCL

ḍábà Mákàdzí: he will be given three farms.

(16) *Yágà úvâh-há hkâná ŋ rét-é dá-vlâ-ŋ-tâ-lò*

Yágà farm-PL three PREP half-FOC FUT-give-EXT-CPL.NOM-3INCL

Yágà: three and a half farm one will give (him)

*ḍábà Yágà úvâh-há hkân-é dá-vlâ-ŋ-tâ-lò*

ḍábà Yágà farm-PL three-FOC FUT-give-EXT-CPL.NOM-3INCL

ḍábà Yágà: three farms one will give (him)

*Slàgháyá úvâh-há hēsá ŋ rét-é dá-vlâ-ŋ-tâ-lò*

Slàgháyá farm-PL two PREP half-FOC FUT-give-EXT-CPL.NOM-3INCL

Slàgháyá: two and a half farm one will give (him)

*ḍábà Mákàdzí úvâh-há hkân-é dá-vlâ-ŋ-tâ-lò*

ḍábà Mákàdzí farm-PL three-FOC FUT-give-EXT-CPL.NOM-3INCL

ḍábà Mákàdzí: three farms one will give (him).

## CHAPTER 8

# BOOTSTRAPPING THE DEVELOPMENT OF MORPHOLOGICAL ANALYSERS FOR "DISPERSED" NGUNI LANGUAGES - A LINGUISTIC INVESTIGATION

Sonja E. Bosch

*Department of African Languages, University of South Africa, Pretoria, South Africa*  
boschse@unisa.ac.za

### 1. INTRODUCTION

The focus in this chapter is on a linguistic investigation into the feasibility of bootstrapping the development of morphological analysers for two 'dispersed' Bantu languages<sup>1</sup> namely Zimbabwe Ndebele and Tanzanian Ngoni<sup>2</sup> by using an existing prototype of a Zulu morphological analyser as point of departure. The choice of these two languages for this particular investigation is based on their relatedness to South African Nguni languages, their minority language status in the countries where they are spoken, and also on the fact that they are resource-scarce languages for which relatively little or no technological development has been done.

Roux (2010:1), in his invited talk entitled *Do we need linguistic knowledge for speech technology applications in African languages?* presented at the Second African Language Technology workshop in Malta, addressed

... the sustainability of generating new and applicable knowledge as a prerequisite for development of applications in African languages. Given the efforts of the current generation of scholars in African languages, and the seemingly declining interest in African linguistics, at least within the Southern African context, the question arises how to ensure and maintain growth in this sector.

In this chapter the importance of linguistic knowledge for text applications (and by implication for speech applications<sup>3</sup>) will be demonstrated, in particular within the context of a rule-based approach to morphological analysis which requires the writing of large numbers of grammar-based rules. It should also become evident that due to the dearth of language resources as well as declining linguistic expertise in most African languages, bootstrapping of morphologically related languages, based on existing applications, can go a long way to reduce development time and efforts of building morphological analysers for lesser resourced languages – spoken



by relatively few people – thereby ensuring technological development for such languages as well. Antonsen *et al.* (2010:2788) confirm that

... there is a large potential for reusing grammatical resources for grammar-based parsers. The results show that linguistic methods can be used efficiently and build systems on recycled knowledge instead of starting from scratch when dealing with new tasks.

Morphological analysis is generally regarded as a basic enabling application that facilitates the development of more advanced tools and practical language processing applications, such as tokenising, part-of-speech tagging, parsing and machine translation. This is particularly relevant to languages belonging to the Bantu language family, which present challenges for automatic analysis of word forms because of their complex morphological nature, and scarceness of electronic lexicons (Bosch 2010:5-6). In order to be of practical use, morphological analysis depends on underlying broad-coverage machine-readable lexicons as fundamental resources.

The research question is whether bootstrapping the development of morphological analysers across language boundaries for two ‘dispersed’ Bantu languages, by using an existing prototype of a Zulu morphological analyser, is feasible. In this chapter, the possibilities of extending the work done for the ‘parent’ language Zulu to two further related languages will be investigated from a linguistic perspective. In other words, the preparatory work reported on in this chapter will, if found viable, serve as basis for implementation in the bootstrapping process in a next step.

The body of the paper is structured as follows: in the next section some background will be given on the existing Zulu morphological analyser prototype **ZulMorph** and its use for experimental bootstrapping of the development of broad-coverage finite-state morphological analysers for three Nguni languages, namely Xhosa, Swati and (Southern) Ndebele. Subsequently the feasibility of extending the bootstrapping process to two further languages, namely Zimbabwe Ndebele and Tanzanian Ngoni will be investigated by first providing a background to these two languages, and then presenting a concise comparative morphological study. Finally conclusions regarding the feasibility of the bootstrapping process will be drawn and future work discussed.

## 2. ZULMORPH AND ITS USE FOR EXPERIMENTAL BOOTSTRAPPING

The development of a broad-coverage finite-state morphological analyser prototype for Zulu (**ZulMorph**) is based on the Xerox Finite-State Tools (Beesley & Karttunen 2003) and is reported on in detail in several publications, e.g. Pretorius and Bosch (2003a, 2003b and 2010). The Xerox software tool **lexc** is used to

enumerate the required and essential natural-language lexicon and to model the morphotactic structure of Zulu words in this lexicon. Subsequently **lexc** source files are produced and compiled into a finite-state network, which renders morphotactically well-formed, but rather abstract morphophonemic or lexical strings. The morphophonological (phonological and orthographical) alternations are modelled with the Xerox tool **xfst**. Here the changes (orthographic/spelling) that take place between lexical and surface words when morphemes are combined to form new words/word forms, are described.

Finally, the **lexc** and **xfst** finite-state networks are combined together into a single network, namely a so-called lexical transducer that includes all the morphological information about the language being analysed, and constitutes the computational morphological analyser of the language, in this case the Zulu morphological analyser **ZulMorph**.

The use of **ZulMorph** for experimental bootstrapping of the development of broad-coverage finite-state morphological analysers for Xhosa, Swati and (Southern) Ndebele is discussed extensively in Bosch *et al.* (2008). It is argued that linguistic relatedness of resource-scarce languages may be systematically exploited by means of a bootstrapping process in order to reduce development time and efforts without compromising on accuracy. According to Bosch *et al.* (2008) the bootstrapping process is done in various stages by reusing the core components of the Zulu analyser for the three additional Nguni languages. These core components are:

- a) the *morphotactics* component, representing the Zulu word structure consisting of word roots, affixes for all parts-of-speech (word categories) as well as a description of the valid combinations and orders of these morphemes for Zulu word formation;
- b) the *morphophonological* (phonological and orthographical) *alternations* component, representing the changes that take place between lexical and surface words when morphemes are combined to form new words/word forms.

The bootstrapping approach described in Bosch *et al.* (2008) functions as semi-automatic support to human linguistic expertise that allows linguists to focus their attention on aspects in which the languages differ. Adapting **ZulMorph** to provide for affix variations in the related languages, e.g. the form of morphemes in the ‘closed’ classes, proved to be a trivial implementation matter. However, certain areas in the grammars of individual languages that differ substantially from those applicable to Zulu required custom modelling and were built into the analyser as additional components e.g. the copula construction of (Southern) Ndebele and the formation of the so-called temporal form in Xhosa.

The results of a preliminary evaluation based on the use of parallel test corpora of approximately 7,000 types each for the four languages, indicate that the “high

degree of shared typological properties and formal similarities among the Nguni varieties warrants a modular bootstrapping approach" (Bosch *et al.* 2008:66).

Therefore the feasibility of extending the bootstrapping process to two further languages, namely Zimbabwe Ndebele and Tanzanian Ngoni will be investigated from a linguistic point of view in this chapter. In the next section a background will be provided on Zimbabwe Ndebele and Tanzanian Ngoni as possible candidate languages for an extension of the bootstrapping process.

### **3. ZIMBABWE NDEBELE AND TANZANIAN NGONI**

The two languages that have been identified for investigation of extending work done for Zulu as described in section 2, are Zimbabwe Ndebele and Tanzanian Ngoni, based mainly on their a) conjunctive writing system which correlates with the writing system of Zulu<sup>4</sup>; and b) relatedness to Zulu since both are 'dispersed' Nguni languages spoken by Zulu descendants, although Zimbabwe Ndebele seems more closely related, while Tanzanian Ngoni is purported/assumed to be distantly related due to influences by the surrounding East Central Bantu languages. In the following section some historical, demographic, geographic and linguistic information will be given on these two languages.

#### **3.1. History of Zimbabwe Ndebele and Tanzanian Ngoni**

During the 19<sup>th</sup> century wars that were waged by the Zulu King Shaka, Zulu speakers were dispersed across the southern and south eastern parts of Africa, and in their migration northwards, fugitives imposed the Zulu language on weaker tribes (Poulos and Msimang 1998:3). Chief Soshangane and Chief Zwangendaba led the first migration of fugitives to the north in about 1823 and eventually settled in what is known today as Mozambique. After a disagreement between the two chiefs in 1831, Zwangendaba migrated further north through Malawi, and finally settled in Tanzania. Descendants of Zwangendaba's followers are still to be found in Malawi, Tanzania and Zambia. The language spoken by them is known as Tanzanian Ngoni and carries the ISO code [ngo]<sup>5</sup>.

The Ndebele Empire, on the other hand, was founded by Chief Mzilikazi after he and his handful of followers had fled the reign of King Shaka in 1827. They eventually settled in the southern part of present day Zimbabwe near Bulawayo between 1838 and 1840. Mzilikazi continued to expand the Ndebele nation by raiding and subjugating neighbouring tribes. The descendants of Mzilikazi's followers speak a Zulu variety called Zimbabwe Ndebele which has the ISO code [nde].

### 3.2. Classification of Zimbabwe Ndebele and Tanzanian Ngoni

According to the new updated Guthrie classification of Bantu languages list (Nurse and Philippson 2003:648), Zimbabwe Ndebele (S44) is classified as a member of the Nguni group (S40) while Tanzanian Ngoni (N12) is classified as a member of the Manda group (N10), as illustrated in the following table:

**Table 1:** Extract of Guthrie’s classification of Bantu language groups S40 and N10

<b>S40 Nguni group</b>	<ul style="list-style-type: none"> <li>•S41 Xhosa</li> <li>•S42 Zulu</li> <li>•S43 Swati</li> <li>•S44 Ndebele (Zimbabwe)</li> </ul>
<b>N10 Manda group</b>	<ul style="list-style-type: none"> <li>•N11 Manda</li> <li>•N12 Ngoni (Tanzania)</li> <li>•N121 Ngoni (Malawi)</li> </ul>

This classification is in contrast to Doke’s (1967:23) earlier classification which still regards Tanzanian Ngoni as belonging to the Zulu cluster of the Nguni group. It is interesting to note that Doke (1967:90) identifies three main Nguni clusters, namely Zulu, Xhosa and Tekeza, with the Zulu cluster comprising the following languages:

**Table 2:** Doke’s classification of the Zulu cluster

Nguni Group							
Zulu Cluster							
<b>Zulu(of Zululand)</b>	<b>Zulu (of Natal)</b>	Lala	Qwabe	Ndebele (of Transvaal)	<b>Ndebele (of Rhodesia)</b>	Ngoni (of Nyasaland)	<b>Ngoni (of Tanganyika)</b>

The rationale behind Doke’s (1967:23) classification is that “Zulu has flung dialects far and wide” and therefore the two Ngoni varieties are regarded as offshoots of Zulu. Precisely this rationale gave rise to the question whether a Nguni language that is classified on the one hand by Doke as belonging to the Nguni group, and on the other hand is classified by Guthrie as belonging to a completely different language group (N10), could also be included in a Nguni (S40) bootstrapping exercise.

### 3.3. Demographics and Geographical Distribution of Zimbabwe Ndebele and Tanzanian Ngoni

Zimbabwe Ndebele is a language of Zimbabwe spoken in the Matabeleland, Bulawayo area by approximately 1,550,000 speakers (Ethnologue 2005). The language is also spoken in Botswana and Zambia with a total of 1,572,800 speakers in all countries. Although Shona, the dominant language of Zimbabwe, has influenced the vocabulary of Zimbabwe Ndebele, the latter is grammatically still essentially a form of Zulu.

Like most Bantu languages, Zimbabwe Ndebele is also considered a resource-scarce language, compared to other world languages, implying that linguistic resources such as large annotated corpora and machine readable lexicons are not available, and the academic, as well as commercial interest to develop them is limited. There are no exhaustive linguistic descriptions, although a number of descriptive grammars were published, e.g. O'Neil (1969), Pelling and Pelling (1974) and Khumalo (2003), while there are numerous unpublished theses such as Hadebe (2002) especially in the field of lexicography, as well as dictionaries such as *Isichazamazwi SeSiNdebele* (Hadebe 2001) a monolingual dictionary.

Electronic resources available for Zimbabwe Ndebele are Pelling's (1971) dictionary that is available on-line for research purposes, and others that centre mainly around corpus resources from the African Languages Lexicon Project (The ALLEX Project 2003) which includes an unannotated on-line Ndebele Corpus (691,268 words) for corpus searches.

Tanzanian Ngoni is a language spoken in southern Tanzania in the Songea district and the Ruvuma region by approximately 170,000 people. The language is influenced by many indigenous languages that are spoken in the areas where the immigrants settled (Ngonyani 2003:1), and therefore it is not surprising that contemporary Ngoni is not regarded as closely related to the Nguni languages of Southern Africa. Miti (1996:85) refers to the low lexical similarity with Zulu, and Nurse (1985:210) even states that Tanzanian Ngoni "is an indigenous southern Tanzanian language, has been for a long time, and has little to do with S.40 Nguni apart from the transferred name."

Tanzanian Ngoni can be regarded as an extremely resource-scarce language since there are only a handful of published linguistic studies on the description and analysis of this language. Over and above the published descriptive grammars of Moser (1983) and Ngonyani (2003), most references are only in unpublished or manuscript form, e.g. a German-Ngoni dictionary, a grammar by Ebner and an unpublished thesis of Mapunda (cf. Ngonyani 2003:100), while only portions of Bible translations dating back to 1891 to 1898 are available (according to Ethnologue 2005). Furthermore, no electronic text corpora are available (to the

author’s knowledge). The dearth of resources is not surprising seeing that Tanzanian Ngoni is a minority language that shows a strong influence of surrounding East Central Bantu languages such as the macro language Swahili which is the main medium of instruction in Tanzania. Ngonyani (2003:4) states that Ngoni is “very restricted in terms of use. Older people still use it in families and among themselves” and that Swahili words are used to replace a large part of the Ngoni vocabulary. The documentation of the language, especially modelling of the grammar for future generations is therefore of the utmost importance.

#### **4. COMPARATIVE MORPHOLOGICAL STUDY**

The feasibility of extending the bootstrapping process as described earlier will be investigated in this section by presenting a concise comparative morphological study between Zulu and Zimbabwe Ndebele on the one hand, and Zulu and Tanzanian Ngoni on the other. It should be emphasised that the research in this chapter focuses on specific linguistic aspects, which were also successfully addressed in the bootstrapping process of Xhosa, Swati and (Southern) Ndebele. In this chapter, the same approach will be followed (for the linguistic preparation) towards determining whether the morphological structures of Zimbabwe Ndebele and Tanzanian Ngoni lend themselves to a similar bootstrapping process. In a next stage the linguistic differences highlighted in this chapter, will form the basis of the implementation of the bootstrapping process and evaluation based on the computational morphological analysis of texts in the two languages under investigation.

The linguistic preparation for implementation of bootstrapping is based on the information contained in the core components for each language. A summary of the core components of **ZuIMorph** as explained in section 2 is as follows:

**Table 3:** Core components of **ZulMorph**

<b>Morphotactics</b>	<b>Affixes</b> for all parts-of-speech (e.g. SC, OC, CL PREF, V SUFF, N SUFF, TAM morphemes etc.)  <b>Pronouns</b> (e.g. absolute, demonstrative, quantitative)  <b>Demonstrative copulatives</b>
	<b>Word roots</b> (e.g. nouns, verbs, relatives, adjectives, ideophones, conjunctions)
	Rules for legal <b>combinations</b> and <b>orders</b> of morphemes (e.g. <i>si-ya-ba-thum-el-a</i> and not <i>*si-ba-ya-thum-a-el</i> )
<b>Morpho- phonological alternations</b>	Rules that determine the form of each morpheme (e.g. <i>ku-lob-w-a</i> > <i>ku-lotsh-w-a</i> , <i>u-mu-lomo</i> > <i>u-m-lomo</i> )

In the following section the equivalent components for each language will be discussed and compared with Zulu (as base language in the proposed bootstrapping process). It should be noted that in this preparatory phase of the bootstrapping process, human intervention in the form of linguistic insights proves to be rather intensive in order to maintain grammatical accuracy (cf. Bosch 2010:11).

**4.1. Zimbabwe Ndebele - Morphotactics Component**

4.1.1. The fixed morphological structure of words in a language is modelled by affixes that form the ‘closed’ morpheme classes to which typically no new items can be added. In cases where the Zulu ‘closed’ morpheme information (morphotactics) cannot be shared with Zimbabwe Ndebele, additional information for the latter language has to be provided as illustrated in Table 4.

**Table 4:** Examples of variations in Zulu and Zimbabwe Ndebele ‘closed’ morpheme information

Morphemes	Zulu	Zimbabwe Ndebele
<b>Subject concords:</b>		
2PL 2PL PST	<i>ni-</i> <i>na-</i>	<i>li-</i> <i>la-</i>
<b>Object concords:</b>		
2PL	<i>ni-</i>	<i>li-</i>
<b>Quantitative pronouns:</b>		
2PL	<i>nodwa</i> <i>nonke</i>	<i>lodwa</i> <i>lonke</i>
<b>Absolute pronouns:</b>		
2PL	<i>nina</i>	<i>lina</i>
<b>Demonstrative pronouns:</b> for position 1 the final vowel of the demonstrative is elided. Conjunctive writing of demonstrative pronoun in [nde] is also noted.		
	<i>lo mfana</i> “this boy” <i>laba bantu</i> “these people” <i>le mifula</i> “these rivers”	<i>l-umfana</i> “this boy” <i>lab-abantu</i> “these people” <i>l-imifula</i> “these rivers”
<b>Relative concords:</b>		
2PL	<i>eni-</i>	<i>eli-</i>
<b>Adjective concords:</b>		
2PL	<i>eniba-</i>	<i>eliba-</i>
<b>Demonstrative copulatives:</b> CL7, CL8, CL10 and CL15 differ between [zul] and [nde], while an optional morpheme <i>-na</i> may be suffixed to [nde] position 3 of demonstrative copulative of all classes.		
CL7	Pos. 1 <i>nasi</i> Pos. 2 <i>nanso</i> Pos. 3 <i>nasiya</i>	Pos. 1 <i>nansi</i> Pos. 2 <i>nanso</i> Pos. 3 <i>nansiya/na</i>
CL8 and CL10	Pos. 1 <i>nazi</i> Pos. 2 <i>nazo</i> Pos. 3 <i>naziya</i>	Pos. 1 <i>nanzi</i> Pos. 2 <i>nanziyo</i> Pos. 3 <i>nanziya/na</i>
CL15	Pos. 1 <i>nakhu/naku</i> Pos. 2 <i>nakho/nako</i> Pos. 3 <i>nakhuya/nakuya</i>	Pos. 1 <i>nanku</i> Pos. 2 <i>nanko</i> Pos. 3 <i>nankuya/na</i>
<b>Connective prefix:</b> in [nde] is <i>la-</i> instead of [zul] <i>na-</i> . Examples:		
	<i>Banamandla</i> “They are strong” > <i>ba-na-amandla</i>	<i>Balamandla</i> “They are strong” > <i>ba-la-amandla</i>



<b>Plural suffix of the imperative:</b> <i>-ni</i> in [zul] as well as in [nde], but the latter has an additional alternative, namely <i>-nini</i> . Examples:		
	<i>Fundani!</i> "Learn" (pl)	<i>Fundani!</i> "Learn" (pl) <i>Fundanini!</i> "Learn" (pl)
<b>Locative prefix with pronouns:</b> CL2 – CL15 absolute pronouns in [zul] use locative prefix <i>ku-</i> while in [nde] they use either <i>ku-</i> or <i>ki-</i> . Examples:		
	<i>kubo</i> "from/to them" <i>kulo</i> "from/to it"	<i>kubo/kibo</i> "from/to them" <i>kulo/kilo</i> "from/to it"
<b>Copulative negative:</b> [zul] noun/pronoun copulatives prefix a negative morpheme; in [nde] noun/pronoun copulatives, positive copula prefix is replaced by negative prefix. Only one of the options of the formation of negative copulative pronouns is similar in the two languages. Examples:		
	<b>Nouns</b> <i>(k)aku-</i> / <i>(k)akusi-</i> or <i>asi-</i> prefixed to copula construction in positive:  <i>(k)aku-ngumuntu</i> <i>(k)akusi-ngumuntu</i> <i>asi-ngumuntu</i> 'It is not a person'  <i>(k)aku -yindoda</i> <i>(k)akusi -yindoda</i> <i>asi -yindoda</i> "It is not a man"	<b>Nouns</b> Nouns beginning with <i>u-</i> > <i>su-</i> : <i>kasu-mfana</i> "He is not a boy"  Nouns beginning with <i>a-/i-</i> > <i>si-</i> : <i>kasi-nkazana</i> "She is not a girl" <i>kawusi-nduna</i> "You are not a chief"
	<b>Pronouns</b> Either <i>(k)aku-</i> is prefixed to copulative of pronoun: <i>(k)aku-yithi</i> "we are not the ones" <i>(k)aku-nguye</i> "he is not the one"  Or <i>(k)akusi-</i> is prefixed to short form of pronoun: <i>(k)akusi-thi</i> "we are not the ones" <i>(k)akusi-ye</i> "he is not the one"	<b>Pronouns</b> In the case of <i>wena/yena</i> , <i>kasu-</i> is prefixed to short form of pronoun: <i>kasu-we/kasu-ye</i> "you/he/she is not the one"  All other pronouns prefix <i>kasi-</i> to short form of pronoun in negative: <i>kasi-mi</i> "I am not the one"

Certain aspects in the Zimbabwe Ndebele grammar require individual modelling, for instance the negative constructions of copulatives derived from nouns and pronouns in Zimbabwe Ndebele that differ considerably from their Zulu counterparts, as well as the orthographic convention of conjunctive treatment of demonstratives. Hadebe (2002:213) lists the rules for spelling and word division in Zimbabwe Ndebele. He states that when the demonstrative pronoun precedes a noun it should be written conjunctively, with the initial of the noun being elided, or the vowel of the pronoun being elided, e.g. *lowomfana* "that boy", *lelolanga* "that

sun", *lababafana* "these boys", *lezizinkomo* "these cattle", *lumfana* "this boy", *lindoda* "this man", *lumfula* "this river". This is in contrast to the case in Zulu, where demonstratives were traditionally written conjunctively with the following noun, until a disjunctive approach was prescribed about 25 years ago, e.g. *leyo nkantolo* "that office", *kulezi zikole* "in these schools".

4.1.2. Word roots in a language form an 'open' morpheme class which contains not only word roots from traditional dictionaries, but is also open to new (derived, compound, coined or borrowed) roots as they become available. Yli-Jyrä (2005:2) emphasises the major challenge of compiling sufficiently extensive and complete word root lexicons (i.e. populating the 'open' word classes), particularly for lesser resourced languages.

A morphological analyser only recognises words of which the roots appear in the underlying lexicon. To make provision for an additional Nguni language, the word root/stem lexicon of **ZuIMorph** will therefore need to be enhanced by the addition of an extensive Zimbabwe Ndebele lexicon. The on-line availability of Pelling's (1971) dictionary (for research purposes) will facilitate automation to a large extent in the next phase when implementation takes place. This dictionary is in HTML format and contains approximately 5,000 entries ( $\pm$  2,000 verbs) with detailed linguistic information.

4.1.3. Word formation rules determine the construction of words/word forms from the inventory of morphemes. There are rules for the combinations and sequences and they are therefore not random, as illustrated in Table 3 above. The rules for legal combinations and orders of morphemes (continuation classes) are identical for Zulu and Zimbabwe Ndebele and therefore can be shared effortlessly.

## 4.2. Zimbabwe Ndebele - Morphophonological Alternations Component

Most alternations occur in instances of palatalisation, which involves a sound change whereby a bilabial or alveolar sound is replaced by a palatal sound in passive formation, locativisation and diminutive formation. Differences in morphophonological alternations between Zulu and Zimbabwe Ndebele are shown in Table 5. All other Zulu alternations treated in **ZuIMorph** apply to Zimbabwe Ndebele.

**Table 5:** Examples of variations in Zulu and Zimbabwe Ndebele morphophonology

Zulu	Zimbabwe Ndebele
Palatalisation with passive, diminutive & locative formation:  <b>th &gt; sh</b> <i>isikhathi</i> > <i>isikhashana</i> "short while"  <b>ph &gt; sh</b> <i>-boph-w-a</i> > <i>wabosh-w-a</i> "he was arrested" <i>iphaphu-ana</i> > <i>iphash-ana</i> "little lung" <i>iphaphu</i> > <i>ephasheni</i> "on/in/at the lung"	Palatalisation with passive, diminutive & locative formation:  <b>th &gt; tsh</b> <i>isikhathi</i> > <i>isikhatshana</i> "short while"  <b>p/ph &gt; tsh</b> <i>-boph-w-a</i> > <i>wabotsh-w-a</i> "he was arrested" <i>impuphu-ana</i> > <i>imputsh-ana</i> "a little mealie meal" <i>impuphu</i> > <i>emputshini</i> "in/at/on the mealie meal"

### 4.3. Tanzanian Ngoni - Morphotactics Component

4.3.1. As in the case of Zimbabwe Ndebele, instances where the Zulu 'closed' morpheme information (morphotactics) does not coincide with or does not exist in Tanzanian Ngoni, additional information for the latter language has to be provided for the bootstrapping process as illustrated in Table 6. In the next stage, decisions will be taken as to the treatment of such variations since some might be more intricate to implement than others.

**Table 6:** Examples of variations in Zulu and Tanzanian Ngoni 'closed' morpheme information

Class	Zulu	Tanzanian Ngoni	Zulu	Tanzanian Ngoni	Zulu	Tanzanian Ngoni
Noun Class Prefixes - no pre-prefixes in Tanzanian Ngoni			Subject concords		Object concords	
1PS			<i>ngi-</i>	<i>ni-</i>	<i>-ngi-</i>	<i>-ni-</i>
1PP			<i>si-</i>	<i>ta-</i>	<i>-si-</i>	<i>-ti-</i>
2PS			<i>u-</i>	<i>u-</i>	<i>-ku-</i>	<i>-ku-</i>
2PP			<i>ni-</i>	<i>m-</i>	<i>-ni-</i>	<i>-va-</i>
CL1	<i>umu-/um-</i>	<i>mu-/m-/Ø-</i>	<i>u-</i>	<i>a-</i>	<i>-m-</i>	<i>-m-</i>

CL2	ba-	va-	ba-	va-	-ba-	-va-
CL1a	u-	∅	u-	∅	-m-	∅
CL2a	o-	∅	ba-	∅	-ba-	∅
CL3	umu-/um-	m-	u-	u-	-wu-	-u-
CL4	imi-	mi-	i-	i-	-yi-	-i-
CL5	i(li)-	li-	<u>li-</u>	<u>li-</u>	<u>-li-</u>	<u>-li-</u>
CL6	ama-	ma-	a-	ga-	-wa-	-ga-
CL7	isi-	chi-	si-	chi-	-si-	-chi-
CL8	izi-	vi-	zi-	vi-	-zi-	-vi-
CL9	in-	n-	i-	i-	-yi-	-yi-
CL10	izin-	n-	<u>zi-</u>	<u>zi-</u>	<u>-zi-</u>	<u>-zi-</u>
CL11	ulu-	lu-	<u>lu-</u>	<u>lu-</u>	<u>-lu-</u>	<u>-lu-</u>
CL12	∅	ka-	∅	ka-	∅	-ka-
CL13	∅	tu-	∅	tu-	∅	-tu-
CL14	ubu-	u-	bu-	u-	-bu-	-u-
CL15	uku-	ku-	<u>ku-</u>	<u>ku-</u>	<u>-ku-</u>	<u>-ku-</u>
CL16	pha-	pa-	ku-	pa-	∅	-pa-
CL17	ku-	ku-	ku-	ku-	∅	-ku-
CL18	mu-	mu-	ku-	mu-	∅	-mu-
CL20	∅	gu-	∅	gu-	∅	-gu-
CL21 <sup>6</sup>	∅	li-	∅	li-	∅	-li-
Tense, Aspect, Mood:	Zulu				Tanzanian Ngoni	
Present tense NB: same order of present tense morpheme in both languages	-ya- appears after SC (irrespective of presence of OC): <i>Bayabiza</i> “They are calling” <i>Bayangibiza</i> “They are calling me”				-i- appears after SC (no OC): <i>Vikema</i> “They are calling” -ku- appears between SC and OC: <i>Vakunikema</i> “They are calling me”	
Remote Past tense	Final vowel -a is used: <i>wahamba</i> “you went”				Perfect suffix -ili is used: <i>wahambili</i> “you went”	
Future tense	Future tense is marked in verb: <i>Bazobiza</i> “They will call”				No future tense marker imbedded verb, <i>yati/mwanja</i> indicating future tense, appears as separate word: <i>Yati vikema</i> “They will call”	
Perfect aspect	Perfect suffix -e/-ile is used: <i>Bahambile</i> “they have gone”				Perfect suffix -ili is used: <i>Vahambili</i> “they have gone”	
Verb extensions: used in suffix position directly following verb root as in [zul], variations in form occur:						
	Applicative -el- Causative -is-				Applicative -il-; -el- Causative -ih-; -is-; -eh-; -es- (choice of applic and caus extension is determined by	

	Passive <i>-w-</i> Reciprocal <i>-an-</i>	vowel harmony with verb stem) Passive <i>-iw-</i> Reciprocal <i>-an-</i>
<b>Absolute pronouns</b>		
	Absolute pronouns for all noun classes	1PS <i>nene</i> / <i>ne</i> 1PP <i>tete</i> / <i>te</i> 2PS <i>veve</i> / <i>ve</i> 2PP <i>nyenye</i> / <i>nye</i> 3PS <i>mwene</i> / <i>mwe</i> * 3PP <i>vene</i> * *applicable to all [+human] nouns in all classes, but not as commonly used as those for 1P and 2P
<b>Possessive:</b> Morphemes written conjunctively in [zul] but disjunctively in [nde] need special treatment in implementation phase, e.g. pre-processing		
	Conjunctive orthography: <i>Izinkomo zomalusi</i> (za-umalusi) "cattle of the herder"	Disjunctive orthography: <i>Ngombe za mdimi</i> "cattle of the herder"
<b>Connective prefix:</b> in both [zul] and [ngo] is <i>na-</i> but with different orthographical conventions as shown below:		
	1PS <i>nami</i> "with me" 1PP <i>nathi</i> "with us" 2PS <i>nawe</i> "with you" 2PP <i>nani</i> "with you" CL1 <i>naye</i> "with him/her" CL2 <i>nabo</i> "with them" etc.	1PS <i>na nene</i> "with me" 1PP <i>na tete</i> "with us" 2PS <i>na veve</i> "with you" 2PP <i>nanyenye</i> "with you" 3PS <i>na mwene</i> "with him/her" 3PP <i>na vene</i> "with them" CL3 <i>nawo</i> "with it" CL4 <i>nayo</i> "with them" etc.
<b>Suffix of the imperative:</b>		
	<i>-ni</i> suffixed to IMP PL <i>Fundani!</i> "Learn" (pl)	<i>-yi</i> suffixed for emphasis to IMP <i>Muyidindayi!</i> "Shut them!"

Ø = morpheme does not exist in particular language; underlined morphemes = identical in both languages

The above is a representative sample of the morphology of Tanzanian Ngoni according to Moser (1983) and Ngonyani (2003). A broad outline of the differences to be expected in the proposed bootstrapping process based on Zulu morphology is given. The differences can be categorised as follows:

- Form of morphemes** – except for a few cases (2PS, CL4, CL5), the form of most concordial morphemes such as basic class prefixes, subject and object concords etc. differs from their Zulu counterparts.
- Non-existent morphemes** – future tense morphemes do not occur as part of the verb construct in Tanzanian Ngoni; nouns do not have class pre-

prefixes nor generic locative prefixes and suffixes nor augmentative and diminutive suffixes.

- c) **Additional morphemes** – the range of class prefixes is more extensive in Tanzanian Ngoni than in Zulu since CL12 (plural) and CL13 (singular) express the diminutive, while CL20 expresses derogation and CL21 the augmentative according to Moser (1983:93). CL16, CL17 and CL18 are used productively to form locatives and therefore have their own agreement morphemes, e.g. *panyumba pa-nyambile* "It was pleasant in the field"; *munyumba mu-nyambile* "it was pleasant in the house".
- d) **Disjunctive morphemes** - possessive morphemes and the connective prefix *na-* follow disjunctive orthographical conventions in Tanzanian Ngoni.
- e) **Double class prefixes** - in addition to containing nouns of a diminutive semantic nature, CL12 and CL13 prefixes are also added to other nouns to indicate the diminutive, e.g. *ka-chipula* "little knife", *tu-chipula* "little knives" (original CL7 prefix *chi-* remains). Similarly the CL20 prefix is added to other nouns to indicate derogation, e.g. *gu-mundu* "ridiculous person" (original CL1 prefix *mu-* remains), etc.  
  
The locative prefixes are also prefixed to nouns with their original class prefixes still in place, e.g. *mu-mfuleni* "in the river" (CL18 locative prefix *mu-*).
- f) **Pairing of classes** – differs from Zulu which typically has the singular/plural combinations CL1/2; CL3/4; CL5/6; CL7/8; CL9/10; CL9/6; CL11/10 versus the following additional class combinations in Tanzanian Ngoni: CL11/6; CL20/6. Examples: CL11/6 *lu-woko* – *ma-woko* "hand/arm"; CL20/CL6: *gu-mundu* – *ma-mundu* "big person".

In comparison to Zimbabwe Ndebele morphology, there is more work to be expected in the case of Tanzanian Ngoni mainly because of the influence of other (Bantu) languages which contributed to more fundamental differences than just the forms of morphemes. Certain aspects in the Tanzanian Ngoni grammar will need to be modelled separately, for instance double class prefixes in the case of nouns that express the diminutive by means of class 12 and 13 class prefixes; or that express location by means one of the locative class prefixes (16 – 18). However, word categories and the order of morphemes coincide to a great extent with Zulu.

4.3.2. The word root lexicon of **ZulMorph** will require enhancing by the addition of as many Tanzanian Ngoni word roots as possible. In the absence of a Tanzanian Ngoni dictionary, any available data can be added for the time being, e.g. word lists totalling  $\pm 1000$  items available in Ngonyani (2003), Moser (1983) and Miti (1996). Since no comprehensive lexicon is available for Tanzanian Ngoni and there are also no regular sound changes between Zulu and Tanzanian Ngoni that can be exploited to facilitate a root lexicon, the identification of Tanzanian Ngoni roots would need to rely on Zulu, similar to Southern Ndebele in the previous bootstrapping experiment (cf. Bosch *et al.* 2008:84). However, the low lexicostatistical percentage (21%) between Tanzanian Ngoni and Zulu as indicated by Miti (1996:85) might hinder acceptable success rates.

4.3.3. As mentioned previously, the construction of words/word forms is governed by rules for legal combinations and orders of morphemes as illustrated in Table 3 above. Ngonyani (2003:49) provides a verb template with the order of morphemes in the verb construction of Tanzanian Ngoni which is very closely related to the Zulu order of verb morphemes, and an investigation into the noun construction of Tanzanian Ngoni leads to the same conclusion, with the exception of double class prefixes as discussed under e) above.

#### **4.4. Tanzanian Ngoni - Morphophonological Alternations Component**

Most alternations in Tanzanian Ngoni occur in instances of two vowels in juxtaposition, which cause a sound change such as vowel elision or consonantalisation or in instances of vowel harmonisation where the choice of vowel is determined by vowel harmony with the stem. Examples of differences in morphophonological alternations between Zulu and Tanzanian Ngoni are shown in Table 7.

**Table 7:** Examples of variations in Zulu and Ngoni morphophonology

<b>Zulu:</b>	<b>Ngoni:</b>
<p><b>Preverbal morphemes:</b></p> <p>SC <i>a-/li-/si-/zi</i> + vowel verb &gt; vowel elision</p> <p><i>ba-ala</i> &gt; <i>bala</i> "they refuse"</p> <p><i>li-eqa</i> &gt; <i>leqa</i> "it jumps"</p> <p><i>a-eba</i> &gt; <i>eba</i> "they steal"</p> <p><i>u-/i-</i> + vowel verb &gt; consonantalisation</p> <p><i>u-akha</i> &gt; <i>wakha</i> "he/she builds"</p> <p><i>i-ala</i> &gt; <i>yala</i> "he refuses"</p>	<p><b>Preverbal morphemes:</b></p> <p>SC <i>i-</i> + <i>-i-</i> (pres tense) &gt; <i>i-</i></p> <p><i>i-i-bwela</i> &gt; <i>ibwela</i> "he/she is coming"</p> <p>SC <i>va-</i> + <i>-i-</i> (pres tense) &gt; <i>vi-</i></p> <p><i>va-i-bwela</i> &gt; <i>vibwela</i> "they are coming"</p> <p>PRS <i>ku-</i> + <i>i</i> (REFL) &gt; <i>kwi</i></p> <p><i>ni-ku-i-pera</i> &gt; <i>nikwipera</i> "I give myself"</p>
<p>Full form <i>umu-</i> with monosyllabic noun stems and vowel initial stems, short form <i>um-</i> with polysyllabic noun stems:</p> <p>CL 1 <i>um-akhi</i> "builder", <i>umu-ntu</i> "person", <i>um-ntwana</i> "child"</p> <p>CL 3 <i>umu-thi</i> "tree", <i>um-lilo</i> "fire"</p>	<p>CL1 and CL3: <i>m-</i> is followed by glide -<i>w-</i> when noun stem is vowel-initial, e.g. <i>mundu</i> "person" vs <i>mwana</i> "child"; <i>mtima</i> "heart" vs <i>mwaka</i> "year"</p> <p>CL14: <i>u-</i> is replaced by glide -<i>w-</i> when noun stem is vowel-initial, e.g. <i>ugimbi</i> "beer" vs <i>wuchi</i> "honey"</p>
<p>Applicative <i>-el-</i></p> <p>Causative <i>-is-</i></p>	<p>Applicative <i>-il-</i>; <i>-el-</i></p> <p>Causative <i>-ih-</i>; <i>-is-</i>; <i>-eh-</i>; <i>-es-</i> (choice of applic and caus extension is determined by vowel harmony with verb stem) e.g.</p> <p><i>-kita</i> "do" &gt; <i>-kitila</i> "do for"</p> <p><i>-pera</i> "give" &gt; <i>-perela</i> "give for"</p> <p><i>-yima-</i> "stand" &gt; <i>-yimiha</i> "stand for"</p> <p><i>-heka</i> "laugh" &gt; <i>-hekesa</i> "make laugh"</p> <p>etc.</p>

An exhaustive comparative study of the entire grammar of Tanzanian Ngoni is beyond the scope of this chapter. However the representative sample of differences in grammatical structure as presented in Tables 6 and 7 provides an insight into the



complexity and intensity of the manual work that would be expected in a bootstrapping exercise for Tanzanian Ngoni.

## 5. CONCLUSIONS AND FUTURE WORK

The contribution of this chapter is particularly focused on providing linguistic inputs for the modelling of regular morphophonological and morphotactic phenomena of the two languages under discussion prior to the implementation of the bootstrapping concept.

The systematic comparative morphological study between the base language Zulu and Zimbabwe Ndebele produces fewer distinguishing features than between Zulu and the three South African Nguni languages, i.e. Xhosa, Swati and Southern Ndebele. With the added advantage of the availability of an electronic lexicon, a bootstrapping procedure as reported on in Bosch *et al.* (2008) seems entirely feasible for Zimbabwe Ndebele. A reiteration process is proposed, starting with a small test corpus and incrementally adapting the base **ZulMorph** in order to increase the success rate of Zimbabwe Ndebele analysis. Linguistic idiosyncrasies can be treated as they arise from the test corpora.

In the investigation of Tanzanian Ngoni grammar more fundamental differences were identified than between other Nguni languages involved in the bootstrapping process and the base language, Zulu, so far. The main reasons being the influence of other (Bantu) languages on the form of various morphemes, the addition and varied functions of class prefixes and so forth. Moreover, only limited word lists are available. However, word categories coincide to a great extent and the order of morphemes agrees with that of the other Nguni languages. Bootstrapping is almost certainly a worthwhile starting point, especially in view of the resource-scarceness of the language. A hybrid approach, also involving statistical and machine learning algorithms, could be investigated with the purpose of adding automatically to the linguistic features of the language. It is proposed that bootstrapping proceeds in a stepwise manner and new information regarding dissimilar morphological constructions be incorporated systematically in the morphotactics component, and rules for the morphophonological alternations component also be adapted in a systematic way in order to increase the success rate of analysis.

It is concluded that from a linguistic perspective it is feasible to bootstrap the development of morphological analysers for two 'dispersed' Bantu languages such as Zimbabwe Ndebele and Tanzanian Ngoni by using an existing prototype of a Zulu morphological analyser, **ZulMorph**. Such resource-scarce languages – with relatively few speakers – can benefit considerably, especially with regard to saving time and expensive human linguistic expertise.

## ACKNOWLEDGEMENTS

Acknowledgement is herewith given to the Computational Morphological Analysis project team for their interdisciplinary cooperation, especially Laurette Pretorius who handles the computational aspects of the Zulu analyser prototype, and Axel Fleisch who was involved with the linguistic aspects of (Southern) Ndebele during the initial bootstrapping experiment.

## ENDNOTES

1. The term ‘dispersed’ here refers to languages spoken by ethnic groups that were dispersed by the raids of King Shaka in the early nineteenth century, known as Mfecane or Difaqane (Poulos & Msimang 1998:3). See also Nurse and Philippson (2003:32).
2. For ease of reference, Ngoni (Tanzania) and Ndebele (Zimbabwe) (cf. Nurse & Philippson, 2003:648-649), are referred to as Tanzanian Ngoni and Zimbabwe Ndebele respectively in this chapter.
3. Cf. Bosch (2010:13) regarding the use of morphological analysis for modelling and prediction of tonal patterns in two projects led by Justus Roux.
4. The orthographical conventions of the Bantu languages, viz. the conjunctive versus disjunctive writing systems, have direct implications for morphological analysis, POS-tagging etc. (cf. Taljard & Bosch 2006). In the case of disjunctively written languages, a certain amount of pre-processing would be required prior to morphological analysis (cf. Bosch *et al.* 2007). See also Prinsloo and Heid (2006) and Van Rooy and Pretorius (2003) for Sotho-Tswana languages in particular.
5. These codes are assigned to languages by the ISO 639-3 standard (ISO 2007) in order to distinguish one language from other languages with the same or similar names and to identify the names of cross-border languages (Ethnologue 2005). The codes for the South African Nguni languages are as follows: Zulu [zul], Xhosa [xho], Swati [ssw] and Southern Ndebele [nde].
6. Moser (1983:93) identifies a noun class 21 indicating the augmentative.
7. It is interesting to note that in Tanzanian Ngoni the basic noun stem *-mfuleni* “river” closely resembles the noun stem with the same meaning in Zulu after the suffixation of the locative morpheme *-ini*.

## REFERENCES

- Antonsen, L., T. Trosterud & L. Wiecheteck. 2010. Reusing Grammatical Resources for New Languages. *Proceedings of the Seventh Conference on International Language Resources and Evaluation*. Valletta, Malta: European Language Resources Association (ELRA).2782-2789.
- Beesley, K.R. & L. Karttunen. 2003. *Finite State Morphology*. Stanford CA.: CSLI Publications.
- Bosch, S. E. 2010. Rule-Based Morphological Analysis: Shared Challenges, Shared Solutions. *Bantu Languages - Analyses, Description and Theory. East African Languages and Dialects*, edited by K. Legere & C. Thornell. Köln: Rüdiger Köppe Verlag. Volume 20:1-15
- Bosch, S., J. Jones, L. Pretorius & W. Anderson. 2007. Computational Morphological Analysers and Machine-Readable Lexicons for South African Bantu Languages. *Localisation Focus - The International Journal of Localisation* 6(1):22-28.
- Bosch, S., L. Pretorius & A. Fleisch. 2008. Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies* 17(2):66-88.
- Doke, C.M. 1967. *The Southern Bantu Languages*. London: Dawsons of Pall Mall. Ethnologue. 2005.  
[http://www.ethnologue.com/ethno\\_docs/distribution.asp?by=area#1](http://www.ethnologue.com/ethno_docs/distribution.asp?by=area#1)  
Accessed: 30-03-2010.
- Hadebe, S. (Ed.). 2001. *Isichazamazwi SeSiNdebele* (a monolingual Ndebele dictionary). Harare: College Press.
- Hadebe, S. 2002. *The Standardisation of the Ndebele Language through Dictionary-making*. Unpublished PhD University of Zimbabwe & University of Oslo.
- Khumalo, L. 2003. *A General Introduction to Ndebele Grammar*. Book Series No. 37. Cape Town: The Centre for Advanced Studies of African Society.
- Miti, L.M. 1996. Subgrouping of Ngoni varieties within Nguni: a lexicostatistical approach. *South African Journal of African Languages* 16(3):83-93.
- Moser, R. 1983. Aspekte der Kulturgeschichte der Ngoni in der Mkoa wa Ruvuma, Tanzania: Materialien zum Kultur- und Sprachwandel. *Beitraege zur Afrikanistik*, Band 17. Wien: AFRO-PUB.
- Ngonyani, D. 2003. *A Grammar of Chingoni*. Munich: LINCOM GmbH.

- Nurse, D. 1985. Review of Rupert Moser's 1983 *Aspekte der Kulturgeschichte der Ngoni in der Mkoa wa Ruvuma, Tanzania: Materialien zum Kultur-und Sprachwandel*. *Journal of African languages and linguistics* 7(1):207-211.
- Nurse, D. & G. Philippon. 2003. *The Bantu Languages*. London: Routledge.
- O'Neil, J. 1969. *A Grammar of the Sindebele dialect of Zulu*. Bulawayo: Ellis Allen PVT. LTD.
- Pelling, J.N. 1971. *A practical Ndebele dictionary*.  
<http://www.linguistics.berkeley.edu/CBOLD/Data/Ndebele.Pelling.1971.txt>  
Accessed: 30-03-2010.
- Pelling, J. & P. Pelling. 1974. *Lessons in Ndebele*. Salisbury: Longman.
- Poulos, G. & C.T. Msimang. 1998. *A linguistic analysis of Zulu*. Pretoria: Via Afrika.
- Pretorius, L. & S. Bosch 2003a. Finite-State Computational Morphology: An Analyzer Prototype for Zulu. *Machine Translation* 18:195-216.
- Pretorius, L. & S. Bosch 2003b. Computational aids for Zulu natural language processing. *South African Linguistics and Applied Language Studies* 21(4):267-282
- Pretorius L. & S.E. Bosch. 2010. Finite State Morphology of the Nguni Language Cluster: Modelling and Implementation Issues. In *Finite-State Methods and Natural Language Processing 8th International Workshop, FSMNLP 2009*, Pretoria, South Africa, July 21-24, 2009, Revised Selected Papers, *Lecture Notes in Computer Science* Volume 6062/2010:123-130. Berlin/Heidelberg: Springer.
- Prinsloo, D.J. & U. Heid. 2006. Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping. *Proceedings of the Lessed Used Languages and Computer Linguistics Conference (LULCL), 27-28 October 2005*, edited by I. Ties. Bolzano: EURAC Research.97-115.
- Roux, J.C. 2010. Do we need linguistic knowledge for speech technology applications in African languages? *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, edited by G. de Pauw., H.J. Groenewald & G-M. de Schryver. Valletta, Malta: European Language Resources Association (ELRA).
- Taljad, E. & S.E. Bosch. 2006. A Comparison of Approaches to Word Class Tagging: Disjunctively vs. Conjunctively Written Bantu Languages. *Nordic Journal of African Studies* 15(4):428-442.
- The ALLEX Project. 2003.  
<http://www.edd.uio.no/allex/corpus/africanlang.html>.

Accessed: 30-03-2010.

- Van Rooy, B. & R. Pretorius. 2003. A word-class tagset for Setswana. *Southern African Linguistics and Applied Language Studies* 21(4):203-222.
- Yli-Jyrä, A. 2005. Toward a Widely Usable Finite-State Morphology Workbench for Less Studied Languages — Part I: Desiderata. *Nordic Journal of African Studies* 14(4):479-491.

## CHAPTER 9

# FROM 'BELEAGUERED' TO 'EMANCIPATED' ORTHOGRAPHY: THE CASE OF NORTHERN SOTHO

Inge Kosch

*Department of African Languages, University of South Africa, Pretoria, South Africa*  
koschim@unisa.ac.za

### 1. INTRODUCTION

The earliest attempts at documenting Northern Sotho words and phrases hail from the period when the Berlin missionaries arrived in the northern part of the former Transvaal Province in South Africa in the early 1860's. The first people among whom the missionaries started their activities were the Bapedi, who spoke the Sepedi dialect. This dialect belongs to the Bantu language Northern Sotho, also referred to as Sesotho sa Leboa in the vernacular. Northern Sotho encompasses a number of dialects and is the standard written form of these dialects. In 1996 it obtained official status as one of the eleven official languages of South Africa, of which nine are Bantu languages. Due to the fact that the written standard language is mainly based on the Sepedi dialect, Northern Sotho is also popularly called Sepedi.

The first task of the missionaries was to learn the dialect and to reduce it to writing in order to be able to spread the gospel more effectively. The initial material produced was invariably gospel material (translations of Scripture and hymns). Apart from these texts, the missionaries also recorded words, phrases and texts in their missionary reports, church bulletins and in the diaries they kept of their activities at the various mission stations. In their choice of orthographic characters they were informed by various factors ranging from simple transfer from their source language (SL) on the one hand to scientifically motivated choices for symbols on the other. With a few exceptions the Latin alphabet has formed the basis for the writing systems of African languages (Internationales Institut für Afrikanische Sprachen und Kulturen 1930:4), the term 'African' here referring to the indigenous languages spoken on the African continent, thus also including the languages of the Bantu language family. The missionaries tended to write independently of one another, and this invariably resulted in the use of a number of different spellings for the same words or the use of graphemes (consisting of one or more characters) that were already in use for other sounds in other languages. The lack of a standardising body gave rise to what one could call a 'beleaguered'

orthography in more ways than one: Not only was there a multiplication of characters and character combinations, but also a proliferation of diacritics. The orthography that is used today can be regarded as an 'emancipated' one in comparison to the earlier writing systems that carried some unnecessary baggage. This discussion focuses on the evolvement of the Northern Sotho orthography towards simplification over a period of time under the pressures of alignment (standardisation) with the other Sotho languages. For a comprehensive treatment of the development of the orthographic symbols for Northern Sotho, the reader is referred to Esterhuyse (1974).

## 2. EARLY ATTEMPTS AT DEVELOPING AN ORTHOGRAPHY FOR N. SOTHO

Esterhuyse (1975:1) points out that the first person to have documented words in Northern Sotho was Alexander Merensky in his article entitled 'Beiträge zur Geschichte der Bapeli' (*Berliner Missionsberichte* xx:353-358, 1862). However, the foundation for the development of Northern Sotho (Sepedi) as a written language, according to Beyers (1981:697), should be attributed to R.F.G. Trümpelmann who started with the adaptation of Biblical stories and German hymns in the idiom of the language in 1868-69, making use of the alphabet of D. Richard Lepsius.

Another pioneer who greatly impacted Northern Sotho as a written language was Karl Endemann. In 1874 he gave a scientific exposition of a proposed orthography for the Sotho languages in an article entitled 'Mittheilungen über die Sotho-Neger'. In 1876 he produced the first grammar for the Sotho languages (mainly dealing with Sepedi, but also including the other two Sotho languages, Southern Sotho and Tswana) entitled *Versuch einer Grammatik des Sotho* (Endemann, 1876b). In this work Endemann applied his own orthography based on the alphabet of Lepsius, which he extended and adapted to meet the special problems posed by the phonetics of the Bantu languages. He used the same orthography later with slight adaptations in his Sotho-German dictionary (*Wörterbuch der Sotho Sprache*) published in 1911.

The earliest examples of written Northern Sotho by the Berlin missionaries were isolated words that occurred scattered in German texts. Several different orthographies came to be used as missionaries devised their own orthographies to the best of their knowledge. How they perceived and documented the sounds depended on their own skills of perception and linguistic training. This often led to inconsistencies in the writings from one author to the next and even in the writings of one and the same author (Harenberg 2009:4). A case in point is the word for 'king' (*kgošī*), for example, in a number of the Berlin missionary reports of the 1880's. At times it is rendered as *kchoshi*, then again as *kchoschi* or *khoshi*. In Endemann's publications (1876a, 1876b and 1911) *kχoši* is used, while the spelling *kxoši* was recommended at a standardising conference in 1929.

### 3. STANDARDISATION EFFORTS

As early as 1837 W.B. Boyce, in the introduction to James Archbell's *A grammar of the Bechuana language* (1837:xix), voiced his concern about the diverse orthographical systems used for Tswana, a concern that was equally applicable to Northern Sotho:

We want to ascertain the moral statistics of South-eastern Africa, the peculiarities of the dialects spoken, and to prepare the way for their acquisition by grammars and vocabularies, in which *one uniform alphabetical system* should be used (own emphasis).

In his publications of 1876a, 1876b and 1911, Endemann also deplored the inaccurate and indiscriminate use of symbols by other writers in the Sotho languages and pleaded for the standardisation of the orthography. He regarded it as his mission to capture the "unadulterated" form of the language for future generations by means of symbols that would reflect each sound and tonal difference accurately. As mentioned earlier, he based his writing system on the standard alphabet of Lepsius, which in his view was the only useful one for these languages. In the introduction to his grammar, Endemann (1876b: Preface) humbly claims to be the first among Zulu and Sotho grammarians to have grasped and represented the laterals correctly. He states that an orthography that does not accurately reflect the pronunciation is an error that is committed genially in the writing of the Sotho languages without due consideration of the existing evidence. He expressed his amazement that this transgression is committed for 'practical considerations', as if issues such as the avoidance of the corruption of the language and the prevention of incomprehensibility in speech and writing were not practical considerations as well! (Endemann 1876a:87). He expounded his orthography in more detail in the front matter of his dictionary (1911), but the proliferation of diacritics rendered the orthography impractical and poorly adapted to the psychical, pedagogical and economic needs of the society it was supposed to serve.

Although Endemann's proposals were highly acclaimed for their scientific precision and insight by some of his contemporaries like C. Meinhof, his orthography remained largely confined to his own work. He obviously did not reckon with the powerful influence exerted by the writing systems that had already started to become established for the Sotho languages, despite existing anomalies that were perpetuated for the sake of convenience. A radical change of an existing writing tradition does not have great prospects of being generally accepted. As indicated by Nida (1963:30 in Harenberg 2009:20) a simpler orthography, even if inaccurate, can become fostered through frequent use and its arbitrary character seems to increase its endearment to people. Endemann's approach, therefore,



however scientific and detailed, was not seriously considered in later attempts at standardisation.

According to Esterhuysen (1975:1) a clear distinction must be drawn between two periods in the development of the orthography for Northern Sotho, namely the period before October 1929 and the period thereafter. The year 1929 signalled an important watershed in that a single orthography was agreed upon for Northern Sotho by the Transvaal Sotho District Committee (Transvaal Education Department). The orthography was subsequently adopted in 1930 and the orthographic rules were published by Lestrade in the same year in an article entitled 'The practical orthography of Transvaal Sotho' (*Northern Sotho Terminology and Orthography* No.2 1962:11). The period before 1929 was characterised as one of multiformity, while the period thereafter saw the achievement of uniformity and standardisation. Earlier, in 1910, an attempt had been made by the Berlin Mission Society responsible for the Northern and Southern Transvaal to standardise the orthography for Northern Sotho at a joint orthography conference in Johannesburg. Its impact was limited, however, as it had not been held at the initiative of the Union Government (Esterhuysen 1974:80, 81). Hoffmann, who attended the conference as one of the representatives of the Berlin Mission Society of the Northern Transvaal, was one of the few authors who implemented the recommendations in a series of articles published in the *Zeitschrift für Kolonial-sprachen* between 1912 and 1916.

The next important milestone followed in February 1947, when a conference was held at the University of South Africa in Pretoria – the so-called Somerset House Conference – to devise a uniform orthography for the Sotho languages (Esterhuysen 1974:97). A number of meetings followed in the next few years until a standardised orthography for Northern Sotho and Tswana was finally approved by the Transvaal Education Department in 1950. It was proposed that the orthography may also be used for Southern Sotho, but the committee members for the latter devised their own orthography, which was approved by the Bantu Language Board in October 1959 (Esterhuysen 1974:105). In 1967 the writing system for Northern Sotho was again adapted slightly because it was decided to eliminate certain alternative symbols with regard to certain speech sounds (Esterhuysen 1974:xiv). The proposed orthographic and spelling conventions have been updated over a period of time in a series of publications, first issued by the Department of Bantu Education and later by the Department of Education and Training under the title *Northern Sotho Terminology and Orthography*. The first publication appeared in 1957. Since then it has undergone revisions in 1963 (No.2), 1972 (No.3) and 1988 (No.4). More recently, the Pan South African Language Board issued an extract from this publication with the latest orthography and spelling rules, written in Northern Sotho (PanSALB 2007).

#### 4. ORTHOGRAPHIC DESIGN

It is said that the orthographies that were developed for the Bantu languages are closely connected to their phonetic representations, as they have largely been designed according to the principle of 'one sound one symbol' (Doke 1954:45). This means that every effort was made to accord each speech sound or phoneme a different symbol in the practical orthography. 'Orthographic depth' is the term used by Joshi and Aaron (2006:569 in Harenberg 2009:17) to describe the relationship between sounds and their visual representations. This relationship can be either 'deep' or 'shallow'. A 'deep' relationship applies when the correspondence is variable and inconsistent, arbitrary and subject to lexical and morphological influences. In the case of the Bantu languages this relationship is 'shallow' because of a close correspondence between sounds and their representations in the practical orthography.

In the next sub-sections various techniques will be considered that writers resorted to to represent Northern Sotho sounds in the practical orthography.

##### 4.1. Choice of Graphemes Based on Sounds they represent in the SL

Authors initially relied on language transfer to document words in the practical orthography in the Bantu languages. By this is meant that they used the familiar characters in the alphabet of their source language (SL) to represent the sounds in the Bantu languages:

German:

*ä* for the mid-low front vowel [ɛ] as in *Bawānda* (Vhavenda)

*sch* for the voiceless prepalatal fricative [ç] as in *moschate* (mošate 'capital city')

*tsch* for the aspirated prepalatal affricate [tʃh] as in *Botschabelo* (Botšhabelo)

*tz* for the ejected alveolar fricative [tsʰ] as in *tzela* (tsela '(foot)path'), and

*ch* for the voiceless velar fricative [x] as in *bochobe* (bogobe 'porridge').

English:

*sh* for the voiceless prepalatal fricative [ç] as in *shupa* (šupa 'show')

*ch* for the aspirated prepalatal affricate [tʃh] as in *chaba* (tšhaba 'flee')

On the analogy of the above, *c* was used for the unaspirated (ejected) version of the prepalatal affricate, i.e. [tʃʰ] as in *cea* (tšea 'take').

It will be noted that the differently inspired orthographies led to some anomalies, for example the letters *ch*, as pronounced in German *Buch* ‘book’, were used to represent the velar fricative [x] in German-based spelling, whereas the same combination, as pronounced in English *church*, was used to represent the prepalatal affricate [tʃh] in English-based spelling. At the orthography conference of 1910 in Johannesburg it had been proposed that the letters *ch* and *c* be used for the aspirated and unaspirated affricates [tʃh] and [tʃʰ] respectively. Although this decision was actually applicable to the Tswana orthography, it had been suggested that these letters also be used for Northern Sotho in a quest to standardise the languages.

Sounds and sound combinations that did not occur in the source languages obviously presented a greater challenge, e.g. the lateral fricative [ɬ] and the labio-prepalatal fricative [β̞ʒ]. As will be indicated later, some of the sounds were subject to quite a number of changes in their practical representation in the course of time until the currently employed orthography was agreed upon.

**4.2. Adoption of Phonetic Representation in the Practical Orthography**

At the conference in 1929 it was decided that the voiced and voiceless velar fricatives in Northern Sotho ([ɣ] and [x] respectively) should be represented by ‘x’ in the practical orthography. This choice seems to have been taken based on the phonetic representation [x] of the voiceless variant. This decision was made independently of other languages or other language groups, as no consideration was given to the fact that the symbol ‘x’ had already become fostered as the symbol for the lateral click in the practical orthography of the Nguni languages. Since Northern Sotho does not have such a click in its sound inventory (except by way of foreign words and loanwords), and since the symbol ‘x’ was still available in the alphabet, the choice fell on this symbol for the velar fricatives [ɣ] and [x]. The velar affricate was similarly affected, i.e. it was written as ‘kx’. However, only a few authors made use of ‘x’ and ‘kx’. The use of these symbols was short-lived as a decision was made to revert back to ‘g’ and ‘kg’ respectively during a revision in 1950:

1910	1929	1950	
gape	xape	gape	‘again’
dikgomo	dikxomo	dikgomo	‘cattle’

### 4.3. Use of Greek Symbols

Where the Latin alphabet was inadequate to represent certain sounds and sound variations, some letters of the Greek alphabet (e.g.  $\chi$ ,  $\upsilon$ ) were introduced. This was largely the practice of one missionary, namely K. Endemann. At the conference in 1910 it had been decided to remove Greek symbols and to replace them with Latin symbols. Nevertheless, in an article that appeared in the 1927-1928 volume of the *Zeitschrift für Eingeborenen Sprachen* (Endemann & Hoffmann, 1927-1928), the Greek symbols were still used. This was because Endemann's son, Chr. Endemann, who was the compiler of the particular article that consisted of a collection of riddles by his father and Hoffmann, still made use of the same orthography that had been proposed by his father. The Greek symbols also appeared in combination with diacritics or characters from the Latin alphabet, e.g.

- $\chi$ : for the voiceless velar fricative [x] or its voiced variant [ɣ], as in *leχapo* 'watermelon' (*legapu*).
- $\chi\iota$ : for the voiceless lateral fricative [ɬ], as in *fiχla* 'arrive' (*fihla*).
- $\check{\chi}$ : for the voiceless prevelar fricative [ʔ] as in *χōile* 'died' (*hwile*).
- $k\chi$ : for the voiceless (aspirated) velar affricate [kxh], as in *mmakχolo* 'grandmother' (*mmakgolo*).
- $\upsilon$ : for the voiced bilabial fricative [β], as in *valoi* 'witches' (*baloi*).

### 4.4. Use of Diacritics

An abundance of diacritics was introduced, particularly by K. Endemann, to cater for all possible sound nuances, including tonal differences.

The International Institute for African Languages and Cultures (Internationales Institut für Afrikanische Sprachen und Kulturen 1930:4), in its guidelines for the practical writing of African languages, remarked on the undesirability of diacritics. Diacritics should be avoided as far as possible, because they clutter a text and strain the eyes. According to the Institute the prolific use of diacritics constitutes a hindrance and danger for its users. As experience has shown, diacritics are prone to being changed (up to the point of being unrecognisable) and trivialised in writing, because users regard them as cumbersome add-ons and they consequently do not pay attention to their accurate representation. When diacritics are omitted, the danger exists that no discrimination is made between dispensable and indispensable diacritics, leading to potential misunderstanding of the meanings of words. The publication of handwritten texts presented a challenge to printers as they could often not decipher the diacritics correctly, with the result that the printed material often contained inaccuracies.

In the interest of simplification and standardisation of the Sotho languages, a decision was made at the 1910 conference to do away with most of the diacritics.

**5. SIMPLIFICATION OF GRAPHEMES FOR THE SPEECH SOUNDS OF NORTHERN SOTHO**

This section aims to show the reduction in the number of diacritics and a simplification of characters in the orthography of Northern Sotho.

**5.1. Vowels**

Vowels were marked by six diacritics in the period before October 1929 (Esterhuysen 1974:18), namely underlining, a grave accent, a circumflex, a dot below a vowel, a horizontal line above a vowel and the breve diacritic  $\sim$  above a vowel. These diacritics were used variously to mark mid-low vowels, raised (narrowed) vowels, lengthened vowels and semi-vowels.

**5.1.1 Mid-low vowels**

The Sotho languages make use of a seven-vowel system, namely:

Phonetic representation	Description	Practical orthography
[i]	high, front	i
[e]	mid-high, front	e
[ɛ]	mid-low, front	e
[a]	low, central	a
[ɔ]	mid-low, back	o
[o]	mid-high, back	o
[u]	high, back	u

In the practical orthography there is no distinction between mid-high and mid-low vowels. Where authors felt they needed to make a distinction, they used various methods to identify the mid-low vowels, e.g. underlining, an accent sign or a circumflex. These conventions were not applied consistently:

- *Underlining* of mid-low vowels (e.g. Hoffmann 1912-1913):

Front vowel: Mid-low e [ɛ] as in *yena* 'he/she'

Back vowel: Mid-low o [ɔ] as in *moloko* 'family'

- *Grave accent* (˘) on mid-low vowels (e.g. Beyer 1920):

Front vowel: Mid-low è [ɛ] as in *èma* 'stand'

Back vowel: Mid-low ò [ɔ] as in *òma* 'become dry'

According to Beyer, the above practice had been agreed upon by the Orthography Conference held at Johannesburg in February 1910 (Beyer 1920:1).

- *Circumflexes* on mid-low vowels (e.g. Schwellnus 1938:31)

Front vowel: Mid-low ê [ɛ] as in *thšabêla* 'flee towards'

Back vowel: Mid-low ô [ɔ] as in *sehlôpha* 'group'

According to Esterhuyse (1974:19) P.E. Schwellnus was the first and only person before 1929 to have used the circumflex to distinguish the mid-low vowels from the mid-high vowels. It was used inter alia in his *Padišô* series of which the first booklet appeared in 1923.

At the 1929 conference it was decided that only one symbol, namely 'e' should be used for both the mid-high [e] and mid-low [ɛ] front vowels, while only 'o' should be used for both the mid-high [o] and mid-low [ɔ] back vowels. In scientific works, however, such as grammars and dictionaries, circumflexes were to be used to indicate the correct pronunciation. In other non-scientific works, diacritics were only to be used where there could be a misunderstanding between two words which are spelt the same but pronounced differently, e.g. *lema* 'plough' vs. *lêma* 'have large horns' or *noka* 'river' vs. *nôka* 'hip' and where the meaning cannot be derived from the context.

#### 5.1.2 Raised (narrowed) vowels

Only the mid-vowels [e, ɛ, o and ɔ] undergo vowel raising or narrowing in certain environments. To indicate raised vowels a dot was used below the vowel in the practical orthography, i.e. ẹ and ɔ̣. These two vowels catered for the raised front and back mid-vowels respectively, with no further distinction as to whether the vowels were mid-high or mid-low. Vowel raising was not marked consistently, neither is it always obvious why certain vowels were indicated as raised in the

positions they occupy, e.g. *o* in *motho* 'person', *shomela* 'work for' (Hoffmann 1912-1913). (Cf. Kotzé 1989:95-96 for conditions under which vowel-raising takes place)

Today the marking of raised vowels is relegated only to narrow phonetic transcription where the symbol [◌̥] is used below the character.

### 5.1.3 Semi-vowels

The breve diacritic (shaped like the bottom half of a circle: ˇ) was used by some authors to indicate semi-vowels, i.e. *o* and *e*:

*Kχokolōana ēa vaImo* 'the little ball of the spirits' (Endemann & Hoffmann 1927-1928:61).

This distinction was not consistently applied, e.g. *ngwana* 'child' appears with three different spellings in the same journal (*Zeitschrift für Eingeborenen Sprachen* 1927-1928), i.e. *ñōana*, *ñoana* (Endemann & Hoffmann 1927-1928:57) and *ngoana* (Hoffmann 1927-1928:245).

Instead of *o* and *e*, most authors used 'w' [w] and 'y' [j], also referred to as vocalic consonants.

### 5.1.4 Lengthened vowels

A horizontal line above a vowel was used to indicate vowel length. In many words its use was confusing. Some authors used it sparingly whereas others used it profusely with the result that its real purpose became obscured. In 1929 the Transvaal-Sotho sub-committee decided that it was not necessary to indicate length (Esterhuyse 1974:83).

## 5.2. Consonants

Once again, the reader is referred to Esterhuyse (1974) for a comprehensive coverage of the symbols that have been used or proposed for Northern Sotho in the past and the time-frames when specific symbols were prevalent. These symbols were informed by individuals' own informed decisions, by the practices of other writers or by recommendations of standardising bodies. The purpose of this overview is to establish the extent of the reduction that was achieved over a period of time, keeping in mind that some consonants retained the same orthographic representation from the beginning. They have been included in the discussion for the sake of completeness.

### 5.2.1 Vocalic consonants

These have already been covered under the discussion of the semi-vowels in 5.1.3.

### 5.2.2 Syllabic consonants

A horizontal line above a consonant was used by Endemann in his dictionary (1911) to indicate a syllabic consonant, e.g. *moñna* 'man'. In earlier publications he used a small circle under the character to indicate a syllabic consonant, e.g. *lla* 'cry', *ñke ke lire* 'may I do' (1876b). This practice was sporadic. Syllabic consonants are no longer indicated by means of a diacritic today.

### 5.2.3 Affricates

For each affricate listed below the orthographic representations are not necessarily indicated in the chronological order in which they occurred in the development of the orthography. The last character listed in each instance (indicated in bold) is the currently used symbol (although this does not imply that the character may not already have been used at some time before the current orthography was agreed upon).

The voiceless aspirated velar affricate [kxh]: kχ, khχ, kx, **kg**

The voiceless ejected labio-alveolar affricate [psʰ]: pz, **ps**

The voiceless aspirated labio-alveolar affricate [psh]: pz, ps, phz, phs, **psh**

The voiceless ejected labio-prepalatal affricate [pʃʰ]: pž, py, pj, pzh, **pš**

The voiceless aspirated labio-prepalatal affricate [pʃh]: pš, psh, phy, ps, py, **psh**

The voiceless ejected alveolar affricate [tsʰ]: tz, **ts**

The voiceless aspirated alveolar affricate [tsh]: ts, ths, **tsh**

The voiceless ejected prepalatal affricate [tʃʰ]: tž, c, tzh, tj, tsh, ts, **tš**

The voiceless aspirated prepalatal affricate [tʃh]: tš, thš, ch, tsh, ths, ts, ty, **tsh**

### 5.2.4 Plosives

The same manner of presentation is followed as explained in 5.2.3.

The voiceless ejected alveo-lateral plosive [tɬʰ]: ɬ, ɬ, **tl**

The voiceless aspirated alveo-lateral plosive [tɬh]: ɬh, ɬh, tl, thl, 'l, **tlh**



The following plosives have always been represented by the same grapheme throughout the development of the orthography:

The voiceless ejective velar plosive [kʰ]: **k**

The voiceless aspirated velar plosive [kh]: **kh**

The voiceless ejective bilabial plosive [pʰ]: **p**

The voiceless aspirated bilabial plosive [ph]: **ph**

The voiceless ejective alveolar plosive [tʰ]: **t**

The voiceless aspirated alveolar plosive [th]: **th**

#### 5.2.5 Clicks

Clicks are unusual consonants in Northern Sotho and only occur in isolated words. The symbols 'x' and 'c' had been used for the velar fricatives [x]/[χ] and the prepalatal affricate [tʃ] respectively as recommended by the 1929 conference. Since these two symbols were already in use for click sounds in the Nguni languages, Northern Sotho followed suit when the orthography was revised in 1950 and decided to use them only for the lateral click x [ɬ] and the alveolar click c [ɭ]. The prepalatal click 'q' [!] completes the number of clicks, though this sound was not found to have occurred in the Northern Sotho writings during the developmental era of the orthography. Clicks remain foreign to the Northern Sotho sound inventory and hence will not form part of the summary of consonants in Table 1 presented in the conclusion.

#### 5.2.6 Flap (retroflex)

The voiced palato-alveolar retroflex [ɭ]: **ɭ, l, d**

#### 5.2.7 Trill

The voiced apico-alveolar trill [r]: **r**

#### 5.2.8 Fricatives

The voiced bilabial fricative [β]: **v** (in some cases printers used 'v' because 'u' was not available on their printing machines), **b**

The voiced labio-prepalatal fricative [β̟]: **vy, vz, vž, by, bz, bž, vzh, bj**

The voiceless bilabial fricative [ɸ]: *f*, **f**

The voiceless labio-dental fricative [ɸ]: **f**

The voiceless labio-alveolar fricative [ɸs]: *fz*, **fs**

The voiceless labio-prepalatal fricative [ɸj]: *fy*, *fsh*, *fž*, **fs**

The voiceless alveolar fricative [s]: **s**

The voiceless alveo-lateral fricative [ɬ]: ʘ, ɬ, ɬ, ɬ plus a diacritic (i.e. 'l, ʘl, 'l, 'l), **hl**

The voiceless prepalatal fricative [ʃ]: *sh*, *s*, **š**

The voiced prepalatal fricative [ʒ]: *ly*, *ly*, *dy*, *zh*, **j**

The voiceless velar fricative [x]: ɣ, x, **g**

The voiceless prevelar fricative [ʰ]: ǀ, 'h, ǀh, 'h, 'h, 'h, h' (i.e. 'h' with six different combinations with diacritics), **h**

The breathy voiced glottal fricative [ɦ]: **h**

#### 5.2.9 Resonants

The alveolar nasal resonant [n]: **n**

The medio-palatal nasal resonant [ɲ]: *nj*, **ny**

The velar nasal resonant [ŋ]: ɳ, **ng**, **n'**

The bilabial nasal resonant [m]: **m**

The alveo-lateral resonant [l]: **l**

#### 5.3. Tone marking

Before 1929 most writers made use of diacritics to indicate tone, but the signs were not applied consistently or in a uniform manner in the different publications. However, high tone was mostly indicated by means of an acute accent (´), e.g. *maaé* 'eggs', *ditlōó* 'small ground beans' (Hoffmann 1915-1916:33, 125).

Karl Endemann felt it necessary to distinguish between five tones. According to him three of them were "main tones", namely high (´), middle (no diacritic, i.e. this was to be left unmarked) and low (˘). The other two tones were mid-high (ˆ) and mid-low (ˊ). These tone markings were only used fully in his dictionary (1911), being a scientific work that was aimed at laying down the correct pronunciations, e.g.

( ' ):	thàva	'mountain'
( , ):	tauḁ	'matter, issue'
( + ):	mo-tzàna	'little village'
( + ):	ṽta	'hide, rot, be mouldy'

Due to the simultaneous use of other diacritics, the added critics for tone often resulted in a concatenation of up to three diacritics on the same symbol.

Writers used their own combinations of diacritics within their chosen systems to distinguish between up to three different tones (high, mid and low). In total eight different symbols were used. Overlaps were bound to occur between the different systems (cf. Esterhuysen 1974 for more detail regarding the individual authors who used these diacritics):

High: ( ' ), ( ' ), ( + )

Mid-high: ( + )

Mid: ( ' ), ( + )

Mid-low: ( + )

Low: ( ` ), ( , ), ( + )

The district committee for Transvaal-Sotho decided in 1929 it was not necessary to indicate tonal distinctions. It was recommended that only two diacritics be retained, namely the acute accent sign ( ' ) for high tone and the grave accent sign for low tone ( ` ), and that these be used only in the case of identical words with different meanings where the correct meaning could not be derived from the context. This means that the eight different diacritic signs to indicate tonal differences that were in circulation among different authors were reduced to only two.

## 6. CONCLUSION

In the beginning of the development of the practical orthography for Northern Sotho there were no combined efforts to standardise the orthography. Individuals were informed by their past training in Europe or observed other scholars' suggestions in this regard, accepting some approaches and rejecting others. The rendering of non-European languages in writing revealed the inadequacy of the Latin alphabet, with the result that writers resorted to other measures such as the introduction of letters from the Greek alphabet and diacritic signs in their quest to

find solutions to the problems. The name of Karl Endemann stands out particularly as a researcher who spearheaded the development of a writing system for the Sotho languages. However, his orthography lacked a participatory and consultative approach and was too scientific to be of any practical use. It was only to be expected that with time practical considerations would supersede scientific ones, even in a scientific work such as a dictionary.

Efforts to standardise the orthography of the Sotho languages since 1910, led to greater uniformity in the choice of symbols. The challenge that the standardising bodies were faced with was to devise a system whereby the same symbols or combinations of symbols would represent the same sounds across the three Sotho languages, as far as practically possible. The decision to do away with all diacritics, where such omission does not cause confusion, was a significant step towards simplification taken by the Somerset House Conference in 1947. Today the only compulsory diacritic is the inverted circumflex used for *š*, *pš*, *pšh*, *tš*, *tšh* and *fš*. The acute and grave accent signs to indicate high and low tones, as well as the circumflexes to indicate mid-low vowels are optional, but they should be used in technical and scientific works or where the correct meaning or pronunciation is not clear from the context.

Vowels have remained relatively uniform in their representations throughout the development of the orthography, but the greatest changes are observed among the consonants. The findings for the consonants are hence presented in the following table, followed by an explanation of the data. The graphemes that are part of the standardised orthography are highlighted in bold in the left column for easy recognition:

**Table 1: Reduction of graphemes for Northern Sotho consonants**

Type of consonant	Total number of graphemes	Graphemes discarded	Graphemes in standardised orthography
<b>Vocalic consonants</b> <i>ö, w</i> <i>ě, y</i>	4	2	2
<b>Affricates</b> <i>kχ, khχ, kx, kg, pz,</i> <i>ps, phz, phs, psh,</i> <i>pž, py, pj, pzh, pš,</i>	29	20	9

phy, <b>psh</b> , tz, <b>ts</b> , ths, <b>tsh</b> , tž, c, tzh, tj, <b>tš</b> , thš, ch, ty, <b>tšh</b>			
<b>Plosives</b> ɔt, t, <b>tl</b> , ʊh, t̥h, thl, 'l, <b>tlh</b> , <b>k</b> , <b>kh</b> , <b>p</b> , <b>ph</b> , t, <b>th</b>	14	6	8
<b>Flap</b> l, l, <b>d</b>	3	2	1
<b>Trill</b> <b>r</b>	1	0	1
<b>Fricatives</b> v, <b>b</b> , vy, vz, vž, by, bz, bž, vzh, <b>bj</b> , f, <b>f</b> , fz, <b>fs</b> , fy, fsh, fž, <b>fš</b> , s, sh, <b>š</b> , x, x, <b>g</b> , 'h, ʔh, 'h, 'h, 'h, h', <b>h</b> , ʃ, xʃ, s, 'l, ʃl, 'l, 'l, <b>hl</b> , ly, ly, dy, zh, <b>j</b>	44	33	11
<b>Resonants</b> <b>n</b> , nj, <b>ny</b> , ñ, <b>ng</b> , m, l	7	2	5
<b>TOTALS</b>	102	65	37

The figures in the second and third columns are not to be construed as indicating the total number of *different* graphemes, the reason being that there are some overlaps across the types of consonants, with the result that some graphemes have been counted twice as it were, e.g. 'l' was used for the flap as well as the lateral resonant. The last column does not include any duplications, but here it needs to be pointed out that there is not in all cases a one-to-one correspondence between the graphemes and the sounds they represent, e.g. 'h' is used both for the voiceless prevelar fricative [h̥] as well as the breathy voiced glottal fricative [ɦ]. There are also some symbols that represent phonetic variants (e.g. 'f' stands for the labiodental fricative [f] as well as the bilabial fricative [ɸ]). This issue was not taken into consideration here, as these variations are encountered on a sub-phonological level.

The totals in the table were arrived at based on the preceding information, but they are not to be construed as absolute figures. Further research may lead to slightly different figures, but this would not change the overwhelming evidence of the extent of consolidation and simplification that has taken place towards a standardised, more practical and user-friendly orthography since Northern Sotho was first reduced to writing.

## ENDNOTE

1. Depending on its combination with other sounds and its position in a word, [ŋ] is represented either as a monograph 'n', as in *nku* [ŋk'u] 'sheep' or as a digraph 'ng' as in *mongwadi* [moŋwadi] 'writer' or *mong* [moŋ] 'owner'.

## REFERENCES

- Archbell, J. 1837. *A grammar of the Bechuana language*. Graham's Town: Cape of Good Hope.
- Beyer, G. 1920. *A handbook of the Pedi-Transvaal Suto language. Practical grammar with exercises, phrases, dialogues and vocabularies*. Morija: Morija Sesuto Book Depot.
- Beyers, C.J. (Ed.) 1981. *SA Biografiese Woordeboek Vol IV*. Durban & Pretoria: Butterworth & Kie Bpk.
- Department of Bantu Education. 1957. *Sotho Terminology and Orthography No.1*. Pretoria: Government Printer.
- Department of Bantu Education. 1962. *Northern Sotho Terminology and Orthography No.2*. Pretoria: Government Printer.
- Department of Bantu Education. 1972. *Northern Sotho Terminology and Orthography No.3*. Pretoria: Government Printer.
- Department of Education and Training. 1988. *Northern Sotho Terminology and Orthography No.4*. Pretoria: Government Printer.
- Doke, C.M. 1954. *The Southern Bantu languages*. London, NY: International African Institute, Oxford Univ. Press.
- Endemann, K. 1874. Mittheilungen über die Sotho-Neger. *Zeitschrift für Ethnologie* 6(1):16-66.
- Endemann, K. 1876a. Die Sotho-Neger. *Allgemeine Missions-Zeitschrift* 3:77-94.
- Endemann, K. 1876b. *Versuch einer Grammatik des Sotho*. Berlin: Wilhelm Hertz.

- Endemann, K. 1911. *Wörterbuch der Sotho Sprache*. Vol.VII of *Abhandlungen des Hamburgischen Kolonialinstituts*. Hamburg: L. Friedrichsen & Co.
- Endemann, K. & C. Hoffmann. 1927-1928. Rätsel der Sotho. *Zeitschrift für Eingeborenen Sprachen* 8:55-74.
- Esterhuyse, C.J. 1974. *Die ontwikkeling van die Noord-Sothoskryftaal*. Unpublished MA dissertation. Pretoria: University of Pretoria.
- Esterhuyse, C.J. 1975. Die ontwikkeling van die skryftekenstelsel van Noord-Sotho. *Studies in Bantoetale* 2(1):1-6.
- Harenberg, M.-A. 2009. *The beginnings of Northern Sotho as a written language: Contributions by the Berlin missionaries*. Unpublished MA dissertation. Berlin: Humboldt Universität zu Berlin.
- Hoffmann, C. 1912-1913. Verlöbniß und Heirat bei den Bassutho im Holzbuschgebirge Transvaals. *Zeitschrift für Kolonialsprachen* 3:124-139.
- Hoffmann, C. 1914-1915. Die Mannbarkeitsschule der Bassutho im Holzbuschgebirge Transvaals. *Zeitschrift für Kolonialsprachen* 5:81-112.
- Hoffmann, C. 1915-1916. Märchen und Erzählungen der Eingeborenen in Nord-Transvaal. *Zeitschrift für Kolonialsprachen* 6:28-54, 124-153, 206-243, 285-326.
- Hoffmann, C. 1927-1928. Sotho-Texte aus dem Holzbuschgebirge in Transvaal. *Zeitschrift für Eingeborenen Sprachen* 18:55-74.
- Internationales Institut für Afrikanische Sprachen und Kulturen. 1930. *Richtlinien für die praktische Schreibung Afrikanischer Sprachen* (2<sup>nd</sup> ed.). London: Internationales Institut für Afrikanische Sprachen und Kulturen.
- Kotzé, A.E. 1989. *An introduction to Northern Sotho phonetics and phonology*. Hout Bay: Marius Lubbe Publishers.
- Merensky, A. 1862. Beiträge zur Geschichte der Bapeli. *Berliner Missionsberichte* xx:353-358.
- PanSALB (Pan South African Language Board). 2007. *Melao ya mongwalo le mopeleto ya Sesotho sa Leboa*. University of Limpopo: Sovenga.
- Schwellnus, P.E. 1938. *Padišô III*. Edendale: Berlin Mission.

## CHAPTER 10

### A SURVEY OF BILINGUALISM IN MULTILINGUAL GABON

Ludwine Mabika Mbokou

*Institut de Recherches en Sciences Humaines, Centre National de Recherche Scientifique et  
Technologique, Libreville, Gabon*  
ludy2407@yahoo.com

#### 1. INTRODUCTION

Natural linguistic diversity is prevalent in most of the sub-Saharan African countries. For these countries, bilingualism is a normal requirement for the daily communication (Mabika Mbokou 2011:375). Several research projects have been undertaken in this area in western countries such as the United States of America, the United Kingdom, France and Germany. In some African countries however, especially in western and central Africa, bilingualism is a sensitive issue (Kwenzi Mikala 1988; Idiata 2002, 2005a & 2011).

This chapter reflects on the notions of bilingualism in the multilingual context of Gabon, a French speaking state in central Africa. A commonly held view among the western Africans is that bilingualism only refers to a person who is able to speak two different European languages such as English and French. For example, a person speaking his/her mother tongue (e.g. Yipunu, one of the Gabonese native languages) and French, is not seen as a bilingual person, while if he/she also speaks English in addition to the languages mentioned, he/she will be considered as a bilingual person on the basis of English and French. Yipunu will not be taken into account. Bilingualism is defined as "*having an effective equal control of two languages*" (Matthews 1997:38). Such a definition certainly excludes most of the people in African countries where the knowledge of more than one language (sometimes more than three) is rather the norm. The problem is that there is no "effective and equal" use of the known or acquired languages in most of these language diversity countries of Africa.

Gabon is the case to be studied in this chapter with the aim to show that bilingualism or even multilingualism is a complex concept with various implications. It is herein hypothesised that even if a person masters two or more languages, there is often a situation whereby one language is the dominant language and the other the weaker language. The second hypothesis of this study is that linguistic situation and function might determine the type of bilingualism that is manifested in specific societies and individuals. Linguistic situation and function



should therefore be taken into account in the definition and description of a noted bilingualism circumstance. It is also known that an opposing group of definitions is based on the idea of language use, maintaining that bilingualism starts at the point where a speaker can first produce complete meaningful utterances in both languages (Johnson & Johnson 1998:29). In this study, we will focus on bilingual children of multilingual Gabon and on their language acquisition process.

## **2. THE ISSUE OF BILINGUALISM**

Bilingualism exists as a possession of an individual. It is also possible to talk about bilingualism as a characteristic of a group or community of people. Bilinguals are most often located in groups, communities or in a particular region. Co-existing languages are often in a process of rapid change. Where many languages exist, there is often language shift. This is one of the reasons why, some-how, the complexity of defining the concept of bilingualism is due to the fact that one must decide on how much knowledge of the second language (L2) one must possess. There certainly exists a difference between a person who can master an L2 and the ability of using it. The person might struggle with the speaking proficiency while the writing proficiency is perfect, and vice-versa (Hoffmann 1991; Bloom 2000; Bhatia and Ritchie 2006). One should therefore use the term bilingualism in its broad meaning, and consider degrees or levels of bilingualism. Most of the second language (L2) users control a different range of registers and styles in the two languages. For this reason a person can have a bilingual knowledge while being a monolingual. And yet, when someone uses more than one language, he/she has to internalise the L2 or the third language (L3) in such a way that he/she can use it for different functions and at different levels. Such a person needs to know more than asking for directions, for a cup of tea or saying a few greetings (Baker and Jones 1998; Hovens 2002; Hill 2009).

Moreover, in daily life, bilingualism is more restrictive. When a person is required to be a bilingual at work or whatever situation it may be, he/she is often expected to know the L2 just good enough to be able to produce what is required. Whether the person is fluent or not in the language all proficiencies (speaking, writing, reading and listening) do not matter as long as he/she satisfies the needs of the situation. And more often, the use of the L2 will be restricted to translation and oral conversation, while the writing proficiency is neglected. In that sense, Johnson and Johnson (1998) indicate an illustration that advertisements for bilingual secretaries seem to require an ability to use the second language for professional purposes alone; those for bilingual teachers often require the ability to teach non-English-speaking children rather than the knowledge of a language (Johnson & Johnson 1998:29).

Another aspect of bilingualism is the fact that languages change with the environment in which they are spoken. Therefore, bilingualism is not a sign of any particular achievements, a bilingual person is not two monolinguals in one person, and as Hoffmann (1991:3) puts it, bilingualism arises as a result of contact. Bilingualism and/or multilingualism are not steady states. This is applied to monolinguals. For instance, it has been witnessed in the past few decades that countries such as Canada, Belgium, Finland and Rwanda have gone from being officially monolingual to official bilingual (e.g. Canada and Belgium) or multilingual (e.g. Finland and Rwanda) countries. It therefore appears that bilingualism is a situation nowadays found in many countries around the world (cf. Richards & Schmidt 2002:52). Likewise, because of immigration, mixed marriages, modern tourism, to name a few reasons, people become bilingual at a younger stage. It has been estimated that two-thirds of the world's children grow up in a bilingual environment (Bhatia and Ritchie 2006). We can surely say that in the long run, being a bilingual will be a world norm.

Most sub-Saharan African countries (if not all) are language diversities. The populations of these countries therefore live in a daily multilingual environment. However, in most of these countries the constitution and the language policy do not have an official status to the existing multilingual phenomenon. For most of them, it is the language of the former colonial administration that is both the official and the national language. This is particularly the case in Gabon with French as particularly shown in Ndinga-Koumba-Binza (2005a, 2007 & 2011). In cases where there are several official languages, the most frequently spoken and used will be the ex-colonial language. For these countries, being bilingual or multilingual is just a fact and not a statement (cf. Richards & Schmidt 2002:52). They cannot be seen as multilingual states as Switzerland is, i.e. a federal republic with four official languages (German, French, Italian and Romansh). An exception can be made with regard to Rwanda that started a multilingual policy at the beginning of the year 2000 due to historical reasons.<sup>1</sup>

In the case of Gabon, however, the majority of the population speaks native languages, but not always fully on a daily basis. In fact, French has the privileged position of being the sole official language (used in government, parliament, education, etc. and as sole medium of instruction). This situation appears to be a hindrance to the promotion of the indigenous languages (cf. Ndinga-Koumba-Binza 2004), but it can be turned into an advantage.

All resources found in French provide great data for translation, terminology, lexicography, etc., in the African languages of Gabon. As a result, the education system can be used to carry this data. For example, in 2010, the Gabonese Ministry of Education set up a committee of linguists, lexicographers, teachers and translators to translate the national anthem (originally written in French) into seven

of the native languages. The translated versions are currently being taught for experimentation in a few primary schools of Libreville, the capital city of Gabon, and of Port-Gentil (Gabon's second biggest city). In addition, discussions in parliament still have to come to conclusion in order to determine the extension of the translation project to the Constitution (which is presently available in French only), so as to translate it into native languages.

It should however be noted that the language-in-education policy can affect the development and maintenance of bilingualism or monolingualism at both the individual and the national level. Hoffmann (1991:8) emphasises this point when saying that it has been attested, for instance, that unless the education system takes proper account of the special needs of the children of minority groups, they will not become fully functional in the minority and majority codes. On the other hand, a minority language that finds its way into the school curriculum will enjoy enhanced prestige and this can, in time, positively affect public attitude towards the language concerned and its speakers, who may find it easier to maintain. The following facts are noted in Gabon:

- (i) Each Gabonese language has a very limited number of speakers (Gabon has a population of smaller than a 1.5 million.);
- (ii) In Gabon, French enjoys a highly privileged status being written, read and spoken on a daily basis while the native languages are only spoken languages; and
- (iii) proposals for writing systems for Gabonese native languages have never reached any agreement for implementation (although very few people have a certain reading and/or writing literacy in their mother-tongues).

The education system in this case appears to be a good place to promote these languages and elevate them to official status (Mabika Mbokou 2011; Ndinga-Koumba-Binza 2005b; Nzang-Bie 2001). The South African example whose post-Apartheid constitution includes nine of the indigenous languages among the 11 official languages at national level is a perfect illustration that any African language can become an official language with political willpower. For the situation in Gabon, one of the main challenges would be the exact number of these indigenous languages (cf. Idiata 2005b, 2008 & 2011; and Ndinga-Koumba-Binza 2005a, 2007 & 2010 regarding the debate on the number of Gabonese languages). The latest inventory by Kwenzi Mikala (1998) indicates 62 (including both languages and dialects), which are grouped into 10 language-units on the basis of mutual intelligibility.

To promote the Gabonese native languages, it is possible to elevate the strongest language of each unit to an official language status (cf. Ndinga-Koumba-Binza 2010). It is thus the duty of linguists and language practitioners to provide

suggestions and recommendations to help the Government in devising and implementing reliable and realistic linguistic policies. Such policies should promote and protect native languages. Consequently, a better definition of bilingualism could be given according to its function. This leads us to differentiate degrees of bilingualism. Johnson and Johnson (1998:30) mention a new terminology, which can put all levels of bilingualism under the same heading: the term bilingualism is replaced by “multi-competence”. Thus, a bilingual person is said to be a “multi-competent” person, no matter the level of the language proficiency. This terminology fits well with the statement that a monolingual person could be a person with a bilingual knowledge.

### **3. BILINGUAL CHILDREN IN GABON**

Generally, the practice and alternative use of two languages will be called bilingualism, and the person involved bilingual. Hence, to be bilingual or even multilingual, a person should at least have a good knowledge of both languages and must be able to speak, read, and/or write them alternatively. This is not the case in Gabon where the so-called bilinguals have full proficiency (speaking, writing, reading and listening) for French only. Speakers of Gabonese indigenous languages (including mother-tongue speakers) have perfect speaking and listening proficiency, but have no real reading and writing proficiency in these languages (Mabika Mbokou 2008 & 2009). They may be classified in the category described by Martinet *et al.* (1969) that indicated the symmetrical versus asymmetrical bilingualism dichotomy. In the first case (symmetrical bilingualism), the person is supposed to have an equal mastery of both languages. He/she must be fluent in both languages, and in a sense, stable. However, when one of the languages is lesser known than the other, it will be a situation of asymmetric bilingualism. In this case of asymmetric bilingualism, one can distinguish:

- Passive bilingualism: The less known language is understood without being spoken. This is the case of Gabonese pupils who learn English in high schools.
- Non-receptive bilingualism: The language can be spoken but not fully understood. It is illustrated with Gabonese children having one of the native languages as second language.
- Written bilingualism: An individual has a good reading proficiency of the language (the language is understood when written), but a poor listening proficiency (the language is not understood when spoken). A number of pupils learning a foreign language (e.g. English, German) are in this situation in Gabonese high schools.

- Technical bilingualism: The knowledge of the language is strictly limited to the professional needs. It is the case of high schools teachers of foreign languages (English, German, Italian, Arabic, etc.) in Gabon.

Mabika Mbokou (2008) indicates that one can find two types of bilingualism among Gabonese bilingual children. The two categories are determined by whether, in a bilingual situation, the child learns both languages simultaneously or not. If the child acquires both his/her parents' mother-tongue and French, it is a case of "early bilingualism". If he/she acquires only the parents' native language, the child will only come to know French at school (mostly at primary school). It is a case of "late bilingualism". The linguistic environment is very important in learning a language. It helps the learner to learn and retain quickly, and not to be confused in his/her practice. The words a person who learns a foreign language will hear most frequently will constitute the basis on which one can build his/her vocabulary (lexicon). These words are usually the ones used in a typical conversation, in the daily life. Such words concern greetings, questions about direction and words used when asking for help. They will have priority; they are the words that the learner will encounter most frequently. Children are no exception. The process by which they learn is similar to that of adults when learning a foreign or new language. There are however two other processes for them to learn a new or foreign language. The first one is through the immediate environment of the family, the linguistic context, and the second one is through reading, through the television and through school.

The first process goes along with daily conversations where familiar words abound. They are the most learnt words, and they are easier to be picked up by the children. These are words that the surrounding adults use repetitively. In the second process, which most of the time refers to books, the complex and academic words will be found. The two ways constitute the two registers that lead to the two acts of the linguistic activity of a community. These acts are the production (encoding) and the understanding (decoding) of the language (Bergenholtz *et al.* 1999). A difference must be made between learning a new or foreign language and acquiring a second language. The first normally refers to adults or pupils who have come into contact with a language through education or a formal medium of instruction. Immigrants, women or men in mixed marriages, children of former colonised countries fall into this category. The second type merely concerns children of a mixed marriage, or children who come into contact with two (or more) languages in their early childhood. They do not learn the languages, but both languages are part of their language acquisition process (cf. Hoffmann 1991, Johnson & Johnson 1998).

In Gabon, both categories of children are found. Most of the children living in Libreville will fall into the first category. The situation is such that they will come

to learn either French as a second language, or one of the Gabonese languages as second language, especially in a monolingual home. For children in a mixed marriage, the best known situation will be learning the Gabonese language as a second language. In the rural areas, both categories of children will also be found with the difference that the native language is the dominant one while in the urban areas the dominant language is French. Thus, Gabonese children often have to face more than two languages in their language acquisition process but most of all in their linguistic environment. They are secondary bilinguals since they come to know either French or one of the native languages later in their acquisition process.

#### **4. THE ACQUISITION PROCESS**

The acquisition or the learning process of a second language implies that the first language is well established in the linguistic competence of the speaker. For this reason, the second (late bilingualism) and first (early bilingualism) categories of children seem to be similar because in both cases, we do have one dominant language and one weaker language. However a difference exists in the fact that in the first case, the child may just learn but not yet be able to master the linguistic system of the learnt language, while in the second category, both languages become well established and mastered, although there will still be one that would dominate the other. Second language learning does require certain conditions during the acquisition process. Two major considerations can be drawn from the acquisition process of a second language by Gabonese children (cf. Ndinga-Koumba-Binza 2011; and Idiata 2009):

- In rural areas, the child will first learn a native language (e.g. Yipunu). It is the language spoken at home, in the neighbourhood, in the community and most of the time in the region. This child will come into contact with French only in primary school for some cases, and most of the time in pre-school. At the early stage of childhood, the native language is both first and dominant language. Following the acquisition of French, the native language position as first language will progressively decline the more the child progresses in his/her education and the more he/she steps outside his/her native region for urban areas where French is the only language of social and inter-ethnic communication (Ndinga-Koumba-Binza 2011). The lack of a standardised written form and the fact that the native language is not used at all levels of education as medium of instruction will also be important factors to the weakening of the native language in favour of French (Idiata 2005a, 2008 & 2009; Mabika Mbokou 2009 & 2011).
- In the urban areas, the child has either the native language or French as first language. A number of children were born and raised in monolingual homes,

the home language being the Gabonese native language shared by both parents. In this case, the native language will be the first language because it has been learnt naturally, and French will be the acquired second language. The child may be fluent in both languages, developing them at almost the same time. The native language will still be the first language but will not have the position of dominant language. In fact, French will be more present in the child's daily life due to its position as the sole medium of instruction at school, the only medium of social and inter-ethnic communication and the only language used in all public domains (media, administration, business, etc.). On the other hand, urban children who would have learnt French as a first language and only came into contact occasionally with a native language as a second language (e.g. during a holiday at the grandparents' home in rural areas) are usually from mixed marriages, i.e. where parents are from different ethnic groups thus from different native languages. The parents' medium of communication is most often French. In this case, French remains the first and dominant language of the child. In the situation where the parents' medium of communication is one of the parents' mother-tongue, the child receives it as first language. The finality in this case is not different from the child whose parents have the same native language as mother-tongue.

It should be noted that in countries such as Gabon where people have to learn a second language for practical purposes (work, school, social integration, etc.), the process of learning is rarely formal. It is mostly natural and informal. In other words, it is done through the daily communication needs that the learner encounters. In this regard, Hoffmann (1991:33-55) argues that when a child learns to speak, he/she learns to use language as a means of expression, communication and social contact. The same child is also learning to use language as a tool for understanding and manipulating the world around him/her, i.e. he/she is learning that he/she needs to hear the language from people who surround him/her. In other words, language is an essential ingredient of a child's socialisation process (Hoffmann 1991:34).

It is common knowledge that the familial linguistic environment is an ideal context for language acquisition for young children (2-4 years old). It can also be a good environment for the older children (8-15 years old) to acquire a second language. This has the advantage of being an active process because there might be a rich extralinguistic context to the conversation. The speaker will often have some sensitivity to the extent of the listener's knowledge, and the listener can ask questions (Bloom 2000:192). The vocabulary found in written sources is mostly like the linguistic environment in which children grow up. Bloom (2000:194) reports on a survey conducted on how many words pupils who read at school can learn in a year. The result was that they will read about half a million words a year and be exposed to about 10 000 unknown words per year. To the contrary, children are

often exposed to rare words in the course of casual conversations at home, such as during meal time. This fact can explain why the linguistic context is the most frequent way for children to learn. It provides a platform to learn both simple and complex words. The active function of learning a word can be defined by "*hearing-questioning*" to get more information from the learner's part. Most of the time the language is learnt by hearing words in the context of sentences and use, and this linguistic context is used to figure out what they mean. The simple sentences will then be the ones most commonly heard. They will often not be able to talk using complex sentences and sometimes they will only use nominal sentences and contracted forms. The particular case of Gabon shows that many children have to learn the language of their parents as a second language because of the high number of indigenous languages. This mostly applies to children having parents belonging to different ethnic groups. These children are raised in French and they will come to know their parents' respective languages at a later stage. All these children may become fluent in their parents' languages, but as mentioned earlier, few will be able to read or write in those languages due to the privileged position of French and the lack of standardised writing systems for native languages.

The process of the apprenticeship in those languages (like with every child or adult learning a new language) will be more directed at the semantic level than the syntactic level. They will become familiar with verbs and nouns such as modal verbs, locatives, nominal classes, nominal and verbal pronouns in linguistic context. It will not be long sentences but simple and short ones, plus short affirmation and interjections forms. Thus Gabonese children can be either bilingual or monolingual. The monolingual children are those having French as first and only language. The bilingual ones are those having French and one of the native languages as main languages. This group of children can be divided into two sub-groups according to their first language. The first sub-group will include all the children who have French as first language. The second one will include the children who have a native language as first language.

## **5. BILINGUALISM IN THE EDUCATION SYSTEM**

A few Gabonese private schools and pre-schools offer an initiation to native languages. Curricula found in these schools are mainly concerned with the acquisition of the vocabulary of the languages offered. It appears that some parents expect their children to learn Gabonese languages through these schools rather than through a natural process (Nzang-Bie 2001; and Mabika Mbokou 2011). However, such a process of apprenticeship is impeded by a situation where the pupils have to learn the native languages as foreign languages. The reason is that, for most of them, the parents' language is not the children's mother-tongue, and in some cases, the parents' language does not even exist in the school curriculum.



This applies particularly to children living in Libreville. They were born and are being raised in Libreville. Some children have not ever been in their parents' home village. For these children, some Gabonese languages are no difference from foreign languages such as German, Italian, etc. In this case, there is often a situation of late bilingualism. It is the case, for instance, of learners who come into contact with their parents' language(s) after they have acquired French as mother-tongue. They learn their parents' language(s) as a school subject. In most cases, the first contact with the parents' languages is when the children hear their parents talking with relatives or friends. This can also be seen with children who have parents that have the same mother-tongue but choose to give a French education to their children (Nzang-Bie 2001). The case of late bilingualism is also seen with children who have one of the Gabonese languages as first language. Most of these children are introduced to French in their first year of primary school, or while playing with friends in the neighbourhood. But for this group of children, French will easily replace the initial language. However, the late bilingualism that characterises the Gabonese children is only at an oral level and this situation leads to multilingual classrooms with only one dominant language: French. The pupils are sent into Yipunu classes, or Fang classes, or Ghetsogo classes, etc. on the basis of their parents' language and not on their knowledge of the proposed taught language.

The needs of the pupils learning the Gabonese languages are not the same, although some may have in common the lack of written knowledge of the Gabonese language that they might have previously acquired. Yet, during the proposed lesson, the emphasis is placed on the teaching of vocabulary and the spoken form rather than the written form. Thus, the teaching does not take into account the linguistic competence of the children and the lessons look like classes for tourists (Mabika Mbokou 2011).

## 6. CONCLUSION

This chapter has shown that bilingualism is a relative concept with specific reference to Gabonese children in a multilingual environment. Although the majority of the languages spoken in Gabon are non-official languages, the linguistic atmosphere prevailing qualifies Gabon to be called a language diversity. Every Gabonese child is raised in a bilingual environment as they are introduced to a second and even a third language at either a later or earlier stage of their childhood. As a result, the younger Gabonese generations is divided into two major groups where French and the native languages are wrestling for the initial language position. It is finally French that becomes the child's main language of communication due to various factors such as the language diversity of Gabonese urban areas, the national language policy and the needs of inter-ethnic communication.

Introducing native Gabonese languages in the education system can help to maintain and improve multilingualism through the younger generations. It will provide tools enabling them to read and write in a language they already speak. As stated in the United Nations Human Rights Committee General Comment N°23, subsection 6.1, it is a right for a minority group to maintain its culture, language or religion. Accordingly, positive measures by States may also be necessary to protect the identity of a minority and the rights of its members to enjoy and develop their culture and language. The Gabonese Government should make use of the rich linguistic environment to promote the development of the Gabonese languages.

## ENDNOTE

1. Most countries in the African west coast have the language of their respective ex-colonial masters as official languages. Cameroun has an official bilingual status having both English and French as official languages after the unification of the French Cameroon and the British Southern Cameroon in October 1961. As for Rwanda, it is nowadays a self-proclaimed English-speaking country after Rwandan nationals who grew up in Uganda and Tanzania returned having English as main language. The current Rwandan Constitution indicates English, French, Kinyarwanda and Swahili as official languages of the Republic of Rwanda. English is now the main requirement in Rwandan public institutions and in higher education. In 2011, Equatorial Guinea which is formerly a Spanish colony added French and Portuguese as other official languages together with Spanish. Cf. Ndinga-Koumba-Binza (2006) for an overview of English in French-speaking Africa.

## REFERENCES

- Baker, C. & S.P. Jones. 1998. *Encyclopedia of Bilingualism and Bilingual Education*. Multilingual Matters. Oxford. Oxford University Press.
- Bergenholtz, H., S. Tarp & H.E. Wiegand. 1999. Datendistributionsstrukturen, Makro und Mikrostrukturen in neueren Fachwörterbüchern. *Frachsprachen. Languages for Special Purposes. An international Handbook of Special-Language and terminology research*, edited by Hoffmann, L. et al. Berlin, De Gruyter. 1762-1832.
- Bhatia, T.K. & W.C. Ritchie. 2006. *The Handbook of Bilingualism*. London. Blackwell.
- Bloom, P. 2000. *How Children Learn the Meanings of Words*. Cambridge, Mass: MIT Press.
- Hill, L.B. 2009. The decline of academic bilingualism in South Africa: a case study in *Language Policy* 8(4):327-349.

- Hoffmann, C. 1991. *An Introduction to Bilingualism*. Longman Linguistics Literacy. London.
- Hovens, M. 2002. Bilingual Education in West Africa: Does it work? *International Journal of Bilingual Education and Bilingualism* 5(5):249-266.
- Idiata, D.F. 2002. *Il était une fois les langues gabonaises*. Libreville: Editions Raponda-Walker.
- Idiata, D.F. 2005a. *Francophonie et politiques linguistiques en Afrique noire: essai sur le projet gabonais d'introduction des langues nationales à l'école*. Libreville: La Maison Gabonaise du Livre.
- Idiata, D.F. 2005b. *Les langues du Gabon: Données en vue d'une classification fondée sur le critère d'intercompréhension*. Cape Town: The Centre for Advanced Studies in the African Society.
- Idiata, D.F. 2008. Le français et les langues gabonaises, du partenariat au linguicide: une analyse des données des enfants tirées du contexte de la ville de Libreville. *Revue gabonaise des sciences du langage* 3:85-208.
- Idiata, D.F. 2009. *Langues en danger et langues en voie d'extinction au Gabon*. Paris: L'Harmattan.
- Idiata, D.F. 2011. *Ces belles idées reçues sur les langues du Gabon*. Paris: Les Points sur les i.
- Johnson, K. & H. Johnson. 1998. *Encyclopaedic Dictionary of Applied Linguistics. A Handbook for language Teaching*. London. Blackwell.
- Kwenzi-Mickala, J.T. 1988. L'identification des unités-langues Bantu gabonaises et leur classification interne. *Muntu* 8:54-64.
- Kwenzi-Mickala, J.T. 1998. Parlers du Gabon: classification du 11-12-97. *Les langues du Gabon*, edited by A. Raponda-Walker. Libreville: Les Editions Raponda-Walker.217-221
- Mabika Mbokou, L. 2008. Le Français langue maternelle! *CENARESTInfos* 4:4.
- Mabika Mbokou, L. 2009. *Model of a Yipunu French School Dictionary*. Berlin. VDM Verlag.
- Mabika Mbokou, L. 2011. Regard sur l'introduction des langues vernaculaires dans le système éducatif du Gabon in *Les Ecritures Gabonaises Tome II*, edited by G.N. Mikala & A. Manfoumbi-Mve. Libreville: Editions Odette Maganga.368-387.
- Martinet, A. et alii. 1969. *La linguistique*. Paris. Denoël.

- Matthews, P.H. 1997. *Oxford concise dictionary of linguistics*. Oxford/New York: Oxford University Press.
- Ndinga-Koumba-Binza, H.S. 2004. Le statut socio-politique du français et la promotion des langues locales au Gabon. Paper presented at the 18th International Conference of the Association of French Studies in Southern Africa (AFSSA) held by the French Section of the Department of Modern Foreign Languages, Stellenbosch University, Stellenbosch, South Africa: 8-11 September 2004.
- Ndinga-Koumba-Binza, H.S. 2005a. Considering a lexicographic plan for Gabon within the Gabonese language landscape. *Lexikos* 15:132-150.
- Ndinga-Koumba-Binza, H.S. 2005b. Politique linguistique et éducation au Gabon: un état des lieux. *Journal of Education* 4(1):65-78. Réduit: Mauritius Institute of Education.
- Ndinga-Koumba-Binza, H.S. 2006. English in French-speaking African countries: The case of Gabon. *The study and use of English in Africa*, edited by A.E. Arua, M.M. Bagwasi, T. Sebina and B. Seboni. Newcastle upon Tyne: Cambridge Scholars Publishing. 152-164.
- Ndinga-Koumba-Binza, H.S. 2007. Gabonese language landscape: Survey and perspectives. *South African Journal of African Languages* 27(3):97-116.
- Ndinga-Koumba-Binza, H.S. 2010. Unités-langues et standardisation dans les langues gabonaises. *Ecriture et Standardisation des Langues Gabonaises*, edited by J. Hubert & P.A. Mavoungou. Stellenbosch: SUN Press. 153-178.
- Ndinga-Koumba-Binza, H.S. 2011. From foreign to national: a review of the status of French in Gabon. *Literator* 32(2):135-150.
- Nzang-Bie, Y. 2001. Vers une éducation multilingue au Gabon: première approche. *Revue gabonaise des sciences du langage* 2:17-29.
- Office of the High Commissioner for Human Rights. *General Comment N°23: The rights of minorities (Art.27)*. 08/04/1994. <http://www.unhchr.ch/tbs/doc.nsf/Opendocument> Consulted 27/07/2011.
- Richards, J.C. & R. Schmidt. 2002. *Longman dictionary of language teaching and applied linguistics*. London: Longman/Pearson Education. Third Edition.

---

## **PART 3:**

### LEXICOGRAPHY AND TERMINOLOGY

---

## CHAPTER 11

# TOWARDS A YILUMBU DICTIONARY OF IDIOMATIC PHRASES<sup>1</sup>

Paul Achille Mavoungou

*Département des Sciences du Langage, Université Omar Bongo, Libreville, Gabon*  
moudika2@yahoo.fr

### 1. INTRODUCTION

The aim of this chapter is to introduce my forthcoming dictionary entitled "*Yilumbu Idiomatic Dictionary*" and to describe some of the problems I have encountered and approaches I have developed during my nine years' work on this project. The dictionary is intended to supplement the currently available *Dictionnaire Yilumbu-Français* (Mavoungou and Plumel 2010).

The research reported in this study goes back to 2003, after the completion of my doctoral studies at Stellenbosch University, South Africa. Data for this project are based on fieldwork carried out before 2003 within the framework of my doctoral research work. The data included many words, sentences, songs, stories and idiomatic expressions in Yilumbu<sup>2</sup>.

In this chapter, I will give an account of the methodology and theoretical assumptions of the work before focusing on the type of oral traditions.

### 2.1. METHODOLOGY AND THEORETICAL ASSUMPTIONS

#### 2.1. What is an Idiom?<sup>3</sup>

In French, in order to understand the great wealth of idiomatic expressions in the language, it is important to give a more precise definition of the term "idiome"<sup>2</sup> and "idiotisme". In fact, the term "idiome" is fairly often used in French, while "idiotisme" is rarer. The English language, for example, has only one term where French has two. Both in the French and English traditions, the term "idiom" can be referred to as a polysemous item. When a term has *de facto*, at least two different meanings, it can be regarded as polysemic. This is the case for the word "idiom" referring according to the context to a system of communication used by a community (language, idiolect, dialect, etc.) or a particular language and typical style of expression or construction that cannot be translated literally.

For this second meaning in particular, the French use the word "idiotisme". For the term "idiotisme", most French dictionaries present, according to the language,

different labels, namely: “Gallicism” (in order to define expressions that are typically French), “anglicism” (for usages typical to English), “germanism” (for locutions peculiar to the German language), etc.

I have said above that an “idiom” is a truly polysemous lexical item. The *Petit Robert 1* (1992:957) presents for this term two meanings. The first sense is that of the general language. In this one, “idiom” is defined as “*a particularity typical to a language, idiotisme*”. The second sense is that of the specialised language, namely linguistics in which the word “idiom” is defined as “*the set of expressions of a community that correspond to a specific way of thinking*”. The *Petit Larousse Illustré* (henceforth abbreviated as PLI, in its 1992 (p.530) and 2004 (p.526) editions) adopts a different approach. The PLI only presents one sense for the terms “idiome” and “idiotisme”:

- Idiom: Any system of linguistic communication used by a community (language, dialect, patois, etc.).
- “idiotisme”: Expression or construction typical to a language and impossible to be translated literally (...).

In the English tradition, the *Oxford Advanced Learner’s Dictionary* (2001:592) identifies for the term “idiom” three significations or main meanings:

1. A group of words whose meaning is different from the meanings of the individual words.
2. The kind of language and grammar used by particular people at a particular time or place.
3. The style of writing, music, art, etc. that is typical of a particular person, group, period or place.

This presentation of the *Oxford Advanced Learner’s Dictionary* has the advantage of presenting all meanings of the word in one article. It is understood from all these definitions that idioms are language particular and the typical style of expressions that cannot be translated literally. In other words, the meaning of an idiom cannot be derived from the meaning of its component parts. To all these definitions, it is important to add the following precision that Ntsanwisi (1985:2) gives in *Tsonga Idioms (A Descriptive Study)*: “Idioms are defined by their fixed character as well as by the unpredictability of their meaning”. Ntsanwisi (1985) also proposes a typology of Tsonga idioms, namely:

1. Idioms based on mythology,
2. Idioms based on superstitions and customs,
3. Idioms based on proverbs,

4. Idioms based on wild animals,
5. Idioms based on domestic animals and poultry,
6. Idioms based on ways and habits of the people,
7. Idioms based on the human body,
8. Idioms based on metaphors,
9. Idioms based on figures of speech (metaphor, metonymy, euphemism, etc.).

## 2.2. What is a Collocation?

According to Blumenthal (2005:265), the term *collocation* (a highly controversial concept) has been the subject of intense discussions among lexicographers, semanticists and linguists for many years. The term *collocation* – which is regarded in linguistics as an international technical term – has a polysemous value. Scientific research supposes that the researcher has at his/her disposal theories and methods in order to construct his/her analysis. Therefore, the standard definition of the term collocation for this work is the one that was formulated by Hausmann<sup>4</sup>. Within this framework, the term collocation covers two main meanings, namely: a quantitative meaning and a qualitative meaning. The quantitative importance of collocations is evident in statistical studies of electronic corpora usually comprising millions of words (I think particularly of “the Bank of English, the COBUILD-corpus of present-day English compiled at the University of Birmingham, which at the time the dictionary was completed comprised more than 320 million words”<sup>5</sup>). In a purely statistical sense if not probabilistic, Blumenthal (2005:265-266) writes:

a relationship between two or more words is said to be “collocative” if the frequency of their co-occurrence is higher than what could have been expected on the basis of their overall frequencies in the corpus. The determination of *collocatives* is therefore based not on an absolute frequency, but rather calculated in a ponderous way.<sup>6</sup>

In its qualitative sense collocation is defined by Blumenthal and Hausmann (2006:3) as a “restricted co-occurrence”.

In Blumenthal and Hausmann’s terms (2006:4), collocation can be defined as a phraseological combination encompassing a **base** and a “collocative”. Relationships between *base* and *collocative* are of a rather complex nature. Let us quote *in extenso* what Blumenthal and Hausmann (2006:4) say:

The base is a word (more specifically the meaning of a word, also referred to as “lexia”) that the speaker chooses freely because it can be defined, translated



and learned without the collocative. The collocative is a word (or the meaning of a word) that the speaker selects according to the base because it can be defined, translated and learned without the base.

A bit further, the authors distinguish the collocation (e.g. *to pass an exam*) from three other phraseological constructions, namely the *free combination* (e.g. *a confirmed bachelor*), the *idiomatic or figurative expression* (e.g. *to get on somebody's nerves*) and the *phraseoterm* (e.g. *cardiac arrest* or *traffic lights*). Given their fixed nature and the unpredictability of their meaning, figurative expressions are classified as *fixed combinations* (cf. Hausmann 1984:398, as cited by Netzlaff, 2005:10-11). The focus of the present work will be on those poly-lexical units coded in the language. They are believed to be single units for speakers and they are therefore learned as a group by them. Examples of these occurrences that are regarded as single units in Yilumbu include the following: **mayaanga mayuma**<sup>7</sup> (period of austerity), **ubola na bakaata**<sup>8</sup> (to eagerly serve one's elders), **ubyaala mafutu**<sup>9</sup> (inherit the wife of a relative who has died), etc. Examples of free combinations available in Yilumbu could include the following: **uba na boolu** (to be lazy), **uba na boma** (to fear, to be afraid of), and so on. They will not be taken into account in this work because they are transparent compounds and therefore the user can retrieve the meaning by merely looking at the different constituent parts<sup>10</sup>. However collocations of which the collocative is fairly idiomatic or collocations derived from proverbs have been considered for inclusion in the planned dictionary. One such collocation is **uba na mamani**<sup>11</sup> used figuratively to refer to the act of becoming mature, to behave with dignity. Another one is the expression **biloongu bi ufura** (quack medicines) that is derived from the proverb *Biloongu bi ufura, uyafura mughatsi o ntsubu*<sup>12</sup> (It is easy to lie when there is nobody around to contradict you).

In what follows, I wanted to come back to some aspects of the terminology on phraseology (the notions of maxim, saying, sentence and so on that I had not considered previously) thus integrating them in a much bigger perspective, namely that of paremiology (encompassing tales, epics, ...) which also comes into the picture in the rendering of the idioms presented in this work. After this essay on the terminological systems of both phraseology and paremiology, I will end the chapter with the presentation of a sample of articles.

### 3. TYPE OF ORAL TRADITIONS: A COMPARATIVE TERMINOLOGY OF IDIOMATIC EXPRESSIONS IN ENGLISH AND YILUMBU

#### 3.1. Introduction

There is a variety of oral genres: rhymes, guesses and short stories of a rather playful and entertaining nature (full of puns), proverbs, maxims, tales, epics, legends, etc. If the distinction between apophthegm, "dictum", maxim, proverb and sentence is relatively well established in French and English for example, however it remains a problem in Gabonese languages in general and in Yilumbu in particular. A typology of sentential literature for Gabonese languages has yet to be addressed and data that could be of some help in the process hardly exist. Therefore, the typology proposed here for Yilumbu should be regarded as a mere starting point. In addition, I chose to take traditions established in French or English, for example, as a point of departure. Then I will see to which extent definitions proposed in the French or the English traditions match the terms attested in Yilumbu.

#### ▪ Sententious Literature on the languages of Gabon: a survey

The Latin word *sententia* is used in particular to refer to a decision given by a court<sup>13</sup>. However, the term also refers to a short group of words forming a descriptive expression, especially one that is often used as a moral precept, a maxim. Sentential literature means all studies on sententious statements, namely: sentence, adage, apophthegm, "dictum" or dicton, maxim, proverb, and aphorism.

Among works on oral traditions<sup>14</sup>, there is an abundance of studies on proverbs, tales, epics and guesses. In the field of folk epics, five (5) ethnolinguistic communities of Gabon already have epic studies. These are the Fang (the Mvet epics), the Bisir (the Mulombi epics), the Obamba (the Olende epics), the Nzebi (the Nzebi epics) and the Bapunu (the Mumbwanga epics).

In the field of paremies (from the Greek word *paroima* "instruction" or "proverb") there is a relative abundance of studies on a particular language<sup>15</sup>, but very few studies in which two or more languages are found together. The work of Raponda-Walker (1993a & 1993b) and Kwenzi Mikala (1996) are the exceptions to this state of affairs. Raponda-Walker's contributions (1993a & 1993b) focused on proverbs, mottos, oaths, war cries and guesses in a number of languages of Gabon. The study of Kwenzi-Mikala (1996) deals with the paremies of six (6) ethnolinguistic communities of Gabon, namely: Benga, Mpongwe, Bapunu, Mahongwe, Mwesa and Bakaniji. Kwenzi-Mikala's book entitled *Parémies d'Afrique Centrale (Proverbes et*

*Sentences*) investigates an interesting definitional and theoretical problem for linguists and metalexicographers working in the field of Gabonese languages. As a matter of fact, the author underlines that he has not made a "(...) *distinction between the different categories that compose the sentential literature: adage, apophthegm, 'dictum', maxim and proverb*" (Kwenzi-Mikala 1996:1).

He therefore replaced the original title of his book: *Répertoire des proverbes gabonais* by *Répertoire des parémies gabonaises*. Research conducted so far in the field of idiomatic expressions in Gabon has focused on ethnolinguistic aspects. Given that idioms, in so far as they are capable of teaching us about how people perceive the world in which they live, it appeared to me that they were the most appropriate linguistic elements for a lexicographic investigation. And to reflect on these linguistic units what's more practical than a linguistic dictionary!

#### ▪ Sentence, apophthegm and aphorism

I have defined above, § 2.1., a *sentence* as a group of words expressing a general statement, a moral precept, a maxim. The lexical items *sentence*, *apophthegm* and *aphorism* are often and wrongly regarded as being interchangeable. This confusion is particularly due to the fact that these terms are synonymous. As absolute synonymy is rare if not impossible, there are differences between these lexical items. A sentence is a thought that is expressed both dogmatically and literarily. In French, the term sentence is commonly used to refer to a decision of courts and counsel tribunal.

The terms *apophthegm* and *aphorism* can be regarded as hyponyms<sup>16</sup> of *sentence* (hyperonym or superordinate). *Apophthegm* (from the Greek word *apophthegma* "sentence") is a concise thought. Hence also its memorable character (which is worthy of being preserved in the memory). CIDE (1996:53) defines aphorism as "*a short, cleverly phrased saying which is intended to express a general truth*". In a rather detailed manner, the PLI (2004:74) defines aphorism as a "*sentence which opposes the conciseness of an expression and the richness of a thought, whose aim is less to express a truth than to induce thinking*". In the definitions of the terms *apophthegm* and *aphorism* appears explicitly the idea of conciseness. It is easy to memorise them.

#### ▪ Maxim, adage and dictum

The terms *maxim*, *adage* and *dictum* are also often used in the wrong way (carelessly and without discriminating one term from another). The term *adage* is generally defined broadly as a proverb, but it is in fact an expression of sentential literature of a legal nature. "*The exception proves the rule*" and "*There's one law for the rich and another for the poor*" are well known adages in English and French<sup>17</sup>.

The maxim like the aphorism and the apophthegm are characterised by brevity. A maxim is a short formula which is by nature often critical or ironic: "*That's life*". This maxim teaches us that in life, when something bad or unlucky happened to us we often have to accept it. If maxims are generally anonymous expressions of wisdom representing the values of a community, certain maxims encompass truths written by authors. For example, the following maxim is from La Rochefoucauld: "*Do not do to others what you will not accept from them*"<sup>18</sup>. Popular maxims, those that speak more specifically about daily life or the weather, are usually referred to as dictums. One of the popular maxims is: "*In April you do not cover a yarn*." (cf. Cellard and Dubois 1985:4)<sup>19</sup>

#### ▪ Tale, legend, epic and myth

*Tale* is a literary genre which falls within the narrative type. It is generally a short story of imaginary facts or claimed to be so, which plunges the reader into a confusing universe, different from the real world. Wonderful or fantastic things usually occur in tales. Fairy tales (e.g. *The Sleeping Beauty* or *Little Red Riding Hood*, etc.), traditional tales (written from an oral tradition), etc. have almost always a moral or didactic goal. A *legend* is a traditional story whereby reality is distorted and embellished. It is full of wonderful elements and in some cases it is based on historical facts that have been transformed by popular beliefs. For example: *The legend of Napoleon*; *The legend of Charlemagne*, *The legend of Tarzan*, *The legend of Nyonda Makita*, and so on. A *myth* is a sacred story in the sense that it generally involves gods. It is a figurative or allegorical representation of an idea (for example: *The Creation myth* (the making of the world by God), *The Greek and Roman myths*. *Epic* is a story that generally tells more or less wonderful events of real heroes but often mythologised (i.e. transformed into myth). The most famous epics have told the story of *Soundjata* or the Mandingo epic story (Tamsir Niane 1960), the *Mvet* or the Fang epic story (Ndong Ndoutoume 1970 & 1975), etc.

#### ▪ Collocations and idioms

I have defined above, § 2.2.1., a collocation as a phraseological combination encompassing a **base** and a "**collocative**". CIDE (1996:258) defines collocation as

a word or phrase which is frequently used with another word or phrase, in a way that sounds correct to people who have spoken the language all their lives, but might not be expected from the meaning.

Collocations are therefore groupings of intangible words that should not be confused with a set of frequent words but which can be modified (for instance,

illustrations or citations presented in dictionaries). In the *Petit Robert 1* (1992:775), at the article **feu**, the user is provided with a series of collocations, namely:

- (i) FEU DE JOIE (= feu allumé en signe de réjouissance à l'occasion d'une fête), *feu de la Saint-Jean* (= feu allumé dans un camp de scouts, etc. et autour duquel on se réunit pour chanter, jouer des saynètes),
- (ii) A, AU FEU (*Cuire à feu doux, à grand feu*),
- (iii) COUP DE FEU (= action vive du feu).

For these collocations, Hausmann and Wiegand (1989:349-350) have mentioned that the entry FEU DE JOIE is an example of the use of a context-free subaddressing procedure. The treatment unit "FEU DE JOIE, *feu allumé en signe de réjouissance à l'occasion d'une fête*" encompasses the non lemmatic address FEU DE JOIE and the semantic item *feu allumé en signe de réjouissance à l'occasion d'une fête*.

From a lexicographic point of view, the main semantic difference between collocations and idioms is that idioms are single lexical items whereas collocations are combinations of lexical items. Furthermore, a collocation differs from an idiom by the fact that it is semantically transparent. As already said, a collocation is the relationship between two words or groups of words that often go together and form a common expression. Collocations illustrate the typical combinations in which the lemma occurs as well as its typical use by mother-tongue speakers, and enable the non mother-tongue speaker to use the language's idiomatic phrases (cf. Tomba Moussavou 2007). Lexicographers (Benson *et al.* 1986: ix-xxxv) make a distinction between **grammatical collocations** and **lexical collocations**.

A grammatical collocation is a phrase consisting of a dominant word (a noun, an adjective or a verb) and a preposition or grammatical structure, such as an infinitive or a clause. In contrast, lexical collocations normally do not contain prepositions, infinitives or clauses. For example, many lexical collocations in French or English consist of a verb and a noun. Compared to grammatical collocations, I have observed that lexical collocations that consist of a verb and noun are more frequent in Yilumbu<sup>20</sup>.

#### ▪ Idioms and proverbs

According to Ntsanwisi (1985:2) and as already mentioned, any definition of the term idiom has to take into account two characteristic features viz. its fixed character and its unpredictability of meaning. These fixed phrasal patterns are typical or peculiar of the language being described. With regard to the first characteristic feature, idioms consist of words, which are habitually used together. Their meaning is unpredictable because it cannot be gathered logically from its component parts. Although idioms and proverbs are fixed expressions, idioms are

not as rigid in form as proverbs. The characteristic feature of the proverb lies in the fact that it is both a figurative and didactic expression (Mavoungou 2010a, 2010b). That is the reason why I chose to deal with idioms and proverbs separately<sup>21</sup>.

### 3.2. The Yilumbu Situation

In Yilumbu the terms maxim, adage, "dictum", apophthegm and proverb are not differentiated from one another and they are expressed by the lexemes *n̄:ngù* (in the singular) and *tsín̄:ngù* (for the plural). This term *n̄:ngù* is to be linked with the verb *ùn̄:ngà* "to understand each other, to come together". To give a proverb is said: *ùbúkà n̄:ngù*. The formula *ùbúkà n̄:ngù, ùbíndíyà n̄:ngù àbà:mbà nà yíyùmà* means that in the situation of enunciation, each proverb that is given must be followed by an explanation.

Generally speaking, idioms presented in the proposed dictionary encompass the following:

1. Practical advice, ex. *mùyétù àkúkatànà má:lù mù mí:sù bàbá:là màsùmù* "a woman does not sit with folded-legs before a men's assembly",
2. Particular events transmitted by a more or less ancient tradition, ex. *mùkéli ndí:ngà, mùpúnù ndí:ngà bátù bà dibá:di* "The Akele and the Bapunu, sons of Ndinga (their first father), are a warlike race",
3. Usage and customs, ex. *dìwé:là díbùrà disáfwi:là* "A marriage with children never dies",
4. Idioms based on proverbs, ex. *ngágngà:yù mí:sù* "Your traditional healer is your eyes" (referring to the proverb "eye passes diviner"),
5. Idiom based on a tale, ex. *ùbyá:là màyènà má úkè:là ngé:mbù* "to wait in vain" (referring to a tale between the leopard and the bat),
6. Idiom based on the parts of the body, ex. *útàlè nà mí:sù* "Look with your eyes but do not touch!",
7. Idioms based on more or less elaborated metaphors, ex. *ùsí:mbà dùβé:ngù* "to age" (literally: to catch a stick).

One also notes that most idioms refer to the animal kingdom and to plant life. The majority of idioms in the form of more or less elaborated metaphors are collocations. Regarding tale, legend, epic and myth, there is only one term that describes both the overall and each of these genres of oral literature: *kú:yù*<sup>22</sup>. The term *kú:yù*<sup>23</sup> refers to past events (*má:mbù mábàngà máyùlù*). The collocation *ùlâ:ndà kú:yù*<sup>24</sup> means, to tell a story or to give an example. The locution *síngà:nù/símbà:nù kú:yù* is frequently used by the narrator as a form of invitation and acceptance<sup>25</sup>. Besides, there is another term that is not strictly speaking a synonym of *kú:yù*, but which conveys the same idea: *kú:ngù*. *Kú:ngù* (*tsíkù:ngù*, plural form) means "history,

*story/stories*". It is worth discriminating it from the terms *bíkù:mbù* and *misámù* which refer to "news". If you want to get some piece of news you say: *mbé bíkù:mbù/misámù!*

#### ▪ *Mottos and Guesses*

According to Vansina (1961), mottos are fixed expressions giving the identity of a specific population group, being a family, a clan, a region or a country. They often encompass panegyric elements (expressing praise). Mottos are often cited in a number of circumstances emphasising the characteristics of the group. Several towns, countries, provinces, states, universities and families thus have their mottos. For Gabon, country of the author of these lines, we have: Union – Work – Justice. Among the Balumbu, a number of clans and villages have their identity embodied in mottos<sup>26</sup>. Even better, men and women usually bear nicknames. These ones include a motto of variable length. Kwenzi-Mikala (1990: 117) uses the term « nickname-motto » which he defines as follows:

The nickname-motto recall such value, such behaviour, such particular action, which the bearer can rightfully be proud of. It is a name that is claimed. It is cited from the moment the bearer has received from his father, his grandfather, or his uncle the explanations of the nickname that are referred to as *mikáki* (sg. *mùkákí*).

The Yilumbu expression *útá:ndà kú:mbù* is used in order to give the explanations (*mikáki*) of a nickname that one bears. CIDE (1996:629) defines a guess as an attempt "to give an answer to a particular question when you do not have the facts and so cannot be certain if you are correct". In Yilumbu, guesses are usually referred to as *mákwà nzáŋgə*. A guess work is as follows:

Player A: *Mákwà nzáŋgə?*

Player B: *Nzáŋgə*.

Player A: *Bàŷàtsi bà tá:tə̀ úŋé:βə̀ bó mənì?* (My father's wives are beautiful in the morning?)

Player B: *Máŷàyi mà kóŷə̀ndù*. (Dried leaves of the banana tree)

Games (*bíŋgánə* or *bisáβùlù*) are practiced daily, but also at night in the moonlight. Given that guesses are mainly practiced by children, they are often referred to as *nə:ŋgù tsì mətélì*<sup>27</sup> as opposed to proverbs: *nə:ŋgù tsì mākàlə̀*.

**Table 1:** *Comparative typology of sententious literature in English and in Yilumbu*

English	Yilumbu
Sentence/locution	ndúbùlù (from the verb ütúbà "to speak")
Maxim	nǎ:ŋgù
Adage	
"Dictum" or dicton	
Apophthegm	
Proverb	
Story	kú:ŋgù
News	bíkù:mbù / mìsámù
Nickname	kú:mbù
Motto	mùkákù (kú:mbù)
Guess	mákwà nzǎŋgà / nǎ:ŋgù tsi mǎtèlì
Tale	kú:yù
Legend	
Myth	
Epic	

#### 4. CONCLUSION

The *Yilumbu Idiomatic Dictionary* represents a significant step forward in the presentation of idiomatic material in Gabonese languages. It is, in fact, the first idiomatic dictionary to be published for Yilumbu. I had hoped before the publication of this study to respond to the challenge of Kwenzi Mikala (1996) and familiarise myself with developments in phraseology to be found within the pages of works published by influential scholars (e.g. Benson 1985, Cowie 1998, Hausmann 2007). On theoretical grounds, I hope this work will be welcomed both by lexicographers and other scholars working on idiomatic expressions.



## ACKNOWLEDGEMENTS

My thanks are due to many field informants who gave me much of the information for the dictionary, on which this chapter is based. I wish to express my deep gratitude to the Zentrum für Interdisziplinäre Afrikaforschung (Centre for Interdisciplinary Research on Africa at the Johan Wolfgang Goethe University of Frankfurt, Germany) which offered me the grant to pursue this research. I thank especially Dr. Stefan Schmid who handled the administrative formalities for my stay in Germany. My stay in the Institut für Anglistik und Amerikanistik (Institut for English & American Studies of the University of Erlangen-Nürnberg, Germany) as well as in the Institut für Angewandte Sprachwissenschaft (Institute for Applied Linguistics of the University of Erlangen-Nürnberg, Germany), gave me the time to complete work on the manuscript and the opportunity to benefit from comments given by Prof. Franz Josef Hausmann and Prof. Thomas Herbst. Last but not least, I wish to thank Dr J.C.M.D. du Plessis and Ms T. Harteveld for encouraging me in my Idiomatic Project and making invaluable suggestions on the improvement of this chapter. The remaining inconsistencies and weaknesses are entirely my responsibility. My thanks also go to anonymous reviewers for their valuable comments and suggestions.

## ENDNOTES

1. Parts of the material presented in this chapter are revised forms included in the introduction of the planned dictionary.
2. This material was captured and transcripts of conversations, dialogues and interviews were computerized in the form of a database or lexicographic files.
3. As far as the survey of Gabonese language is concerned, the term "*idiom*" was made known by Raponda-Walker's (1932) article entitled "*L'Alphabet des idiomes gabonais*" which was published in the *Journal de la Société des Africanistes* 3(2):305-314. In the French tradition, "*locution*" is the most commonly used word in order to refer to "*idioms*".
4. I refer specifically to Hausmann (1984:398, as quoted by Netzlaff 2005:10-11), Blumenthal and Hausmann (2006) and Hausmann (2007).
5. cf. Thomas Herbst *et al.*, 2004:vii.
6. All translations from German to English, from French to English and from Yilumbu to English are my own.
7. Literally: "*the streams are dried*".
8. Literally: "*To be wet with elders*".
9. Literally: "*To inherit ancient plantations*".
10. They are treated in the *Dictionnaire Yilumbu-Français* (Mavoungou & Plumel, 2010).
11. Literally: "*To have stones*".

12. Literally: *"if you want to lie about medicine don't you dare to lie to your wife"*.
13. In the *Cambridge International Dictionary of English* (CIDE 1996:1294) the word *sentence* is defined as "a punishment given by a judge in court to a person or organization after they have been found guilty of doing something wrong".
14. Oral traditions refer to a chain of oral and reported testimonies. Oral traditions are thus different from written statements (cf. Vansina 1961:22).
15. See Bodinga-Bwa-Bodinga and Van der Veen (1995) and Adam (1970) for more detailed information.
16. Hyponymy is generally regarded as a relation of inclusion. In fact, the meaning of a lexical item A is the hyponym of the meaning a lexical item B when the extension of A is totally included in the extension of B.
17. The French adages are : « *L'exception confirme la règle* » ; « *Il n'y a pas de règle sans exception* » ; « *Nul n'est censé ignorer la loi* ».
18. The French quotation is : « *Ne fais pas à autrui ce que tu ne veux pas qu'on te fasse* ».
19. The French translation is : « *En avril, ne te découvre pas d'un fil* ».
20. This phenomenon is known in other languages too. Earlier, Tomba Moussavou (2007:238) stressed that lexical collocations that consists of a verb and a noun are more frequent in Yipunu (B43), a sister speech form of Yilumbu.
21. Proverbs will be the object of a separate publication (Mavoungou, forthcoming).
22. The Bapunu use the term *tsáβù* (*dùsáβù* singular form) or *kú:γù* in order to name tales.
23. Literally: « *to chase a story, a tale* ».
24. « *The narrator invites the audience to participate to the performance of the epic and the audience responds positively. When telling the story, the narrator uses again and again this formula in order to sustain the attention of listeners, to indicate the passage from one episode to the other, from one action to the other and after each song* » (cf. Kwenzi-Mikala 1993:117).
25. The motto of *Idúkà mürimà* is *bùdúkà ò mé:sù, dyé:là ó mbùsà*.
26. i.e., « *names that are given to a child together with his birth name or a name that someone chooses when he is grown up or after initiation* » (Kwenzi-Mikala, 1990:117).
27. *Nɛ̀:ngù tsi mǎtéli* and *nɛ̀:ngù tsi mǎkàlǎ* (or *nɛ̀:ngù tsi mǎtsáyǎlǎ* ) literally mean « *proverbs that are said in a standing position* » and « *proverbs that are said in a sitting position* ». In fact, one can play in a standing position (*bisáβùlù bàvási mǎtéli*) but in order to get together and have a serious talk (encompassing the use of proverbs) people usually sit (down) and talk about the problem (*ùsóbà mǎtéli, bàvákálǎ*). However, during palavers, as a sign of politeness you should stand (up) when speaking.

## REFERENCES

### Dictionaries and Encyclopedias

- CIDE: Procter, P. (Ed.).1996. *Cambridge International Dictionary of English*, Cambridge: Cambridge University Press.
- Dictionnaire Hachette Encyclopédique*. 1980, Paris: Hachette.
- OAD: *Oxford Advanced Learner's Dictionary*. 2001. Oxford: Oxford University Press.
- PLI: *Petit Larousse Illustré*. 1992. Paris: Larousse.
- PLI: *Petit Larousse Illustré*. 2004. Paris: Larousse.
- PR : *Petit Robert 1*. 1992. Paris: Robert.
- NPR : *Le Nouveau Petit Robert*. 2008. Paris: Robert.

### Other Literature

- Adam, J.-J. 1970. *Proverbes, devinettes, fable Mbede*. Imprimerie Saint Paul (Issy les Moulineaux) France.
- Benson, M. 1985. Collocations and idioms. *Dictionaries, Lexicography and Language learning*: 61-68. Oxford: Pergamon Press Ltd.
- Benson, M., E. Benson & R. Ilson. 1986. *Lexicographic Description of English*. Amsterdam/Philadelphia: John Benjamins.
- Blumenthal, P. 2005. Le Dictionnaire des collocations: un simple dictionnaire d'exemples? *L'exemple lexicographique dans les dictionnaires français contemporains*, edited by M. Heinz. Tübingen: Max Niemeyer Verlag.265-282.
- Blumenthal, P. & F.J. Hausmann. 2006. Présentation: collocation, corpus, dictionnaire. *Langue Française* 150:3-13.
- Bodinga-bwa-Bodinga, S. & L.J. van der Veen. 1995. *Les proverbes evia et le monde animal (la communauté traditionnelle evia à travers ses expressions proverbiales)*. Paris: L'Harmattan.
- Cellard, J. & G. Dubois. 1985. *Dictons de la pluie et du beau temps*. Paris: Belin.
- Hausmann, F.J. 1984. Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts* 31:395-406.
- Hausmann, F.J. 2007. *Collocations, phraséologie, lexicographie (Etudes 1977-2007 et bibliographie)*. Aachen: Shaker Verlag.

- Hausmann, F.J. & H. E. Wiegand. 1989. Component Parts and Structures of General Monolingual Dictionaries: A Survey. *Wörterbücher.Ein Internationales Handbuch zur Lexikographie/ Dictionaries. An International Encyclopedia of Lexicography/Dictionnaires. Encyclopédie Internationale de Lexicographie*, edited by F.J. Hausmann et al. Berlin: Walter de Gruyter.328-360.
- Herbst, T., D. Heath, I.F. Roe, & D. Götz. 2004. *A Valency Dictionary of English: A corpus-based analysis of the complementation patterns of English verbs, nouns and adjectives*. Berlin/New York: De Gruyter.
- Kwenzi-Mikala, J.T. 1990. L'anthroponymie chez les Bapunu du Sud-Gabon. *Pholia. Revue du laboratoire de Phonétique et Linguistique Africaine* 5:113-120.
- Kwenzi-Mikala, J.T. 1993. La gestualité et les interactions dans la narration d'une épopée: l'exemple de mumbwanga. *Pholia. Revue du laboratoire de Phonétique et Linguistique Africaine* 8:109-120.
- Kwenzi-Mikala, J.T. 1996. *Parémies d'Afrique Centrale (Proverbes et Sentences)*. Libreville: Éditions Raponda-Walker.
- Mavoungou, P.A. 2000. *A Frequency List of the Yilumbu Language*, Unpublished study conducted at the Bureau of the *Woordeboek van die Afrikaanse Taal* (WAT). Stellenbosch: Bureau of the WAT.
- Mavoungou, P.A. 2010a. *A dictionary plan of Yilumbu: Metalexicographical criteria for the compilation of a trilingual dictionary Yilumbu-English-French*. Saarbrücken: VDM Verlag.
- Mavoungou, P.A. 2010b. *Lexicographie et Confection de Dictionnaires au Gabon*. Stellenbosch: SUN Press.
- Mavoungou, P.A. & B. Plumel, 2010. *Dictionnaire Yilumbu-Français*. Libreville: Editions Raponda-Walker.
- Mavoungou, P.A. (forthcoming). *Dictionnaire de 1000 Proverbes Yilumbu (avec des Equivalents en Français et en Anglais)*.
- Mavoungou, P.A. (forthcoming). *Dictionnaire Idiomatique du Yilumbu*.
- Ndong Ndoutoume, T. 1970. *Le Mvett*, livre 1. Paris Présence Africaine.
- Ndong Ndoutoume, T. 1975. *Le Mvett*, livre 2. Paris : Présence Africaine.
- Netzlaff, M., 2005. *La Collocation adjectif adverbe et son traitement lexicographique*, Norderstedt: Herstellung und Verlag.
- Ntsanwisi, H.W.E. 1985. *Tsonga idioms. A descriptive study*, Braamfontein: Sasavona Publishers.

- Raponda-Walker, A. 1932. L'Alphabet des idiomes gabonais. *Journal de la Société des Africanistes* 3(2):305-314.
- Raponda-Walker, A. 1993a. *1500 Proverbes, Devises, Serments, Cris de guerre et Devinettes*, Versailles : Classiques Africains.
- Raponda-Walker, A. 1993b. *3000 Proverbes (multilingue)*. Versailles: Classiques Africains.
- Tamsir Niane, D. 1960. *Soundjata ou l'épopée Mandingue*. Paris: Présence Africaine.
- Tomba Moussavou, F. 2007. *Metalexicographical Criteria for a Monolingual Descriptive Dictionary Presenting the Standard Variety of Yipunu*. Unpublished doctoral dissertation. Stellenbosch: University of Stellenbosch.
- Van der Veen, L.J. 1999. *Les Bantous Eviya (Gabon-B30): Langue et Société traditionnelle*. Notes de synthèse soutenue en vue de l'obtention de l'Habilitation à Diriger des Recherches en Sciences du Langage. Université Lumière-Lyon 2.
- Vansina, J. 1961. *De la tradition orale: essai de méthode historique*. Tervuren: Annales série in 8<sup>e</sup> Sciences Humaines n° 36. MRAC.

## **SAMPLE OF THE PLANNED DICTIONARY ARTICLES**

### **How to Read the Sample Articles?**

Idiomatic expressions presented in this work are alphabetised by keywords (according to the word tradition, lexical items are entered in their complete forms, i.e. prefix plus stem, while in the stem tradition lexemes are lemmatized under the stem without their prefixes (a more detailed presentation of this issue can be found in Van Wyk (1995)). Because there is no fixed word introducing idioms it was decided that they should start with a particular lexical item (either the first noun or the first verb of the idiom) given as a keyword. This keyword will be followed by its part of speech indicator as well as its class number (following conventions used in Bantu languages).

The keyword is then followed by the idiom together with its full phonetic transcription presented between square brackets. The phonetic transcription of the treated idiom is followed by its literal meaning (between inverted commas). Then follows the definition (introduced by the symbol ■) of the idiom as well as illustrative examples (in roman for Yilumbu and in italics for English) presenting the context and co-text in which the idiom typically occurs. This notion of supporting entries (that either belongs to the co-text or the context) is important. The co-text refers to the syntactical environment of the treated idiom whereas the context gives the pragmatic environment of that idiom. Pragmatic information will be introduced by the symbol ⊗.

Extra-linguistic aspects of the treated idiom will be presented in the encyclopedic section of the article and they will be introduced by the black diamond (◆). For collocations in particular, the definition will be followed by a list of the most common collocatives used with the base of the collocation. This list will be introduced by the symbol Ω. As far as illustrative examples are concerned all figurative expressions attested in English will be highlighted. As a general rule they will be taken from CIDE (2006).

## A

**A onomat.**

**atsiyi !** [ätsíyí!]. "He has eaten!". ■ Said about a person who dies at a very advanced age. ⊗ This exclamation, which is used both in a proper sense and figuratively, suggests that the person whom is referred to (usually a father or a mother) has had a wonderful life, he/she was filled with joy by his/her children. Mutu be ghuna mbe ananuuna! **Atsiyi!** *This man was very old when he passed away! He has had a wonderful life!* Ya YIINDU mbe anasiimba duvheengu, mbe anavhembula! **Atsiyi** maboti ma baana! "Yaya" (big brother) MIHINDOU's back was bent with age; he was very old when he died! He was filled with joy by his children.

## B

**BAAKU onomat.**

**na baaku** [nà bá:kù]. [Associative + onomatopoeia] "With (a) word/sound (most probably an onomatopoeic term derived from the verb *ubaku* « to produce a sound, to speak »), but the meaning of the idiom is rather « without a word, speechless »". ■ Unable to speak because you are so angry, shocked or surprised. Ω This idiom is only used with the verb to hear (**ughuulu**). Usomughuulu **na baaku**. *You won't hear him talking.* Ilumbu MUWEEMBI aseenga MUSA AVHU isanamughuulu **na baaku**. *The day MOUWEMBI copiously insulted MOUSSAVOU he stood there without a word.* Vho mughatsi NYUUNDU anwaana na KUUMBA, mi

isanaghuulu NYUUNDU **na baaku**. *When GNOUNDOU's wife and KOUMBA fought each other, NYOUNDOU was speechless with indignation.*

**BAKU onomat.**

**bilu baku bitsi baku** [bílù bákù bítsi bákù].

"The sky + onomatopoeia (expressing a sudden sound e.g. when something is hold tightly), the grounds + onomatopoeia expressing a sudden sound (e.g. when something is hold tightly)". ■ At all cost(s), at any cost. Ω This idiom is often used in threatening formulas. Bumbaatsi yetu naaghu bumana leelu, **bilu baku bitsi baku** uvambe mboongu tsyaami! *I want my money back at any cost, even if it meant giving up our friendship!* **Bilu baku bitsi baku**, leelu tumonasana! *Today, I'll show you that I'm older than you!*

**BAANA n. cl. 2**

**baana ba maghena** [bá:nà bà máyènà]. "The cubs of the leopard". ■ Said about people with whom it is preferable not to have any contact. Mi baana baami **baana ba maghena** duyasiimba! *Beware of my children! Like the cubs of the leopard, it is preferable for you to keep your distance from my children!* Yooghu bana **baana ba maghena**, duyabiila bavhe maloongi! *They are the cubs of the leopard, do not try to get closer to them or even give them some advice!* ◆ Generally speaking, animals that have just given birth are very protective vis-à-vis their young. As far as the leopard (large carnivorous mammal of the cat family) is concerned, it is suicidal to run the risk of attacking or getting closer to its cubs.

**(u)neena baana** [ùné:nà bá:nà] (disapproving)  
 "To defecate children". ■ To be too fertile.  
 Vhana ayendila na NGOMA, kadi muvhu ubura, kadi muvhu ubura: baana avayenaneena! *Since she started going out with NGOMA, each year she is giving birth to a child: she is "defecating" children with this man who has not yet paid the bride price to her parents.*

**baana ba mughuughu** [bá:nà bà mǔyǔ:yù] "The cubs of the coucal". ■ Said about people who cannot go unnoticed. (Prov.) Baana ba mughuughu banawe mbiingi banabasangila, banawe mbiingi banabasangila: **baana ba mughuughu** di di di. *Wherever they go, children of the coucal are recognisable. Yooghu bana baana ba mughuughu nziimbu usaandi. Those are the cubs of the coucal: they are easily recognisable by their bottom which is always wet.* ♦ For local people, the bottom of the coucal is rotten and therefore they believe that eating its meat may cause hemorrhoids.

**BIBEENGA** n. cl. 8

**bibeenga na bipiinda** [bibé:ngè nà bìpî:ndà]. "Pieces of red and black cloths". ■ Financial contribution. Mwaana akaweela sa ik'uroomba **bibeenga na bipiinda**. *Our child has decided to get married; we must now give him our financial contribution.* Atsituula mwa **bibeenga na bipiinda**. *He brought some cash.* Mwaana anabelugha, ika uroomba **bibeenga na bipiinda** mu yefuta divhaana. *Our child has now recovered from his illness; it is time now for us to pay for the services provided by the traditional healer.* Afwaana utuula **bibeenga na bipiinda** mi ikuvhutugha o nzubu. *Before I return home I demand financial compensation, said Kathy to her husband.*

**BIYOGHU** n. cl. 8

**(u)si (mutu) biyoghu** [ùsì mùtù bìyóyù]. "To make noise to someone". ■ To irritate someone. **Uyantsi biyoghu!** *Don't bother me!* Baana bana **bavasi biyoghu!** *These*

*children are disturbing people, they make a great noise!* Lyongu mi itsimaanga botugha vhana, uvasi batu **bighoyu!** *Go away and leave us in peace!* Ami itsituba nana, **uvantsi biyoghu!** *Imosi bí? Yes I am the one who have said, you annoy me! Aren't you glad?*

**BOMA** n. cl. 14

**(u)ba/usiimba na boma** [úbà:ùsí:mbà nà bómà]. "To be caught by fear/to catch fear". ■ To be afraid. **Aba na boma** bu utuba maanga akitsinyoogha. *She was afraid (that) he might be upset if she told him.* Atsisung'usiimba na boma. *He felt suddenly afraid.*

**BOOLU** n. cl. 14

**(u)fu na boolu** [úfù nà bós:lù]. "To die with laziness". ■ To be very lazy. **MAGHANGA** boolu buvamuboka, opodi umana muusu vhana akala, asana isalu. **Ofu na boolu** buna! *MAGANGA is a very lazy person, she can spend the day sitting doing nothing. She will die of laziness!* **MABIKA** ana bitu bivhola, **ofu na boolu**: bambaatsi bakumusalaanga. *MABIKA is very lazy, he relies on others.*

**BUMBAATSI** n. cl. 14

**bumbaatsi bu mulembu na maamba** [bùmbá:tsì bù mùlɛ:mbù nà mǎmbà]. "A friendship between oil and water". ■ Relationship that cannot be trusted, a friendship that is not sincere. Buna **bumbaatsi bu mulembu na maamba**. *I just don't trust him, he always seems very false.* Yooghu ba bwaali **bumbaatsi bu mulembu na maamba**, politiika! *They just pretend to be good friends. That's politics!* ♦ This idiomatic expression criticises false friendships. You should beware of people you don't know thoroughly. As a prosperous person you usually have a lot of friends, but very often they turned out to be false friends and abandoned you when you needed them most. This is why self-interested friendship is referred to as

the relationship between oil and water: oil on top and water below.

**(u)kaanga bumbaatsi** [ùkǎ:ngǎ bũmbá:tsi]. "To build/to tie friendship". ■ To foster friendship. Mupyaapi na iboonga bakaanga bumbaatsi. Vho nziitu, bumbaatsi butsifu toondu dibula. (in fairy tales) *In olden times, the tortoise and the woodcock were very good friends, but their friendship was destroyed by gossip.*

**bumbaatsi bu malamu** [bũmbá:tsi bù málámũ]. "The friendship of alcohol/wine". ■ Said of self-interested friendship. Avamulaanda mu ibili ana mboongu, **bumbaatsi bu malamu!** *His friendship with him is surely motivated by self-interest.*

**BUSINA** *n.* cl. 14

**busina bu matsweela/matsanga** [bũsínǎ bù mǎtswè:lǎ/mǎtsǎngǎ]. "The wealth of the tears." ■ Mocking way of talking about the number of deaths within a particular family. **Busina buna bu matsweela** nooghu nzuumbel! *It seems that they die every day in this family!* Dufwaafu duli na batu bana, piira **busina bu matsweela!** *They just had several deaths in their family: people are dying one after the other; they are really dropping off like flies!*

**busina bu miisu** [bũsínǎ bù mí:sũ]. "The wealth of the eyes." ■ Mocking way of talking about the number of people having large eyes within a particular family. O bwaala bwooghu, bana piira **busina bu miisu!** *I am telling you in their family they really have large eyes!*

**BUTEMU** *n.* (Loan) cl. 14

**butemu bu Corneille** [bũtémũ bù kɔrnej]. "The testimony of Cornelius." ■ False testimony. A! Botugha vhana na **butemu bu Corneille!** *Ah! Go away with your false testimony!* Sa **butemu bu Corneille**, mangeenza! *I am not telling lies, it's the truth!* Ukeba **butemu bu Corneille** asa

muboti! *Be careful you can go to jail if your testimony isn't really true!* Buna **butemu bu Corneille**, MAPIINDA mufitsi. *Don't trust him - APINDA is lying.* ♦ In this idiomatic expression, it is alluded to MIHINDOU [pronounced miyĩ:ndũ] Cornelius (of the Bayengi clan). MIHINDOU Cornelius lived in Dikoundou [pronounced dikũndũ], a village near Tchibanga. Each time he was asked to give the names of the eyewitnesses of his storytelling, he always gave the names of deceased persons. So the expression *butemu bu Corneille* came to be known in the language as synonymous with «false testimony/statement».

**BUTOGHU** *n.* cl. 14

**butoghu buvatobigha** [bũtɔɣũ bũvátɔbíɣǎ]. "The soil is pierced". ■ A word of warning used by parents to make their children aware of the fact the world is full of dangers. This idiomatic expression is derived from the tale about the elephant and the turtle. The elephant - relying on its physical strength - ignored the turtle's warning: "the earth is full of holes" and he ended up by falling into a big hole that was dug by hunters.

## D

**DIBAGHALA/BABAGHALA** *n.* cl. 5/2

**(u)ba dibaghala ufutu kefu** [dibǎɣǎlǎ ùfútũ kǎfũ]. "To be the man of whom the expert in this tradition of circumcision has spitted chilli (onto the penis during the circumcision test)". ■ To be circumcised; to be a man, a real man; a genuine man. ☉ This idiom is generally used in the imperative as a threat. Ke uli **dibaghala ufutu kefu** kwesama! *If you are a man, a real man, come closer!* ♦ In the past during the circumcision ceremony, the expert in this tradition chewed a few seeds of the highly spiced kola/cola (*Buchholzia macrophylla*) that he then spitted onto the



phallus, the circumcised penis. The seeds of *Buchholzia macrophylla* were chosen for the occasion because of their highly spiced or hot nature. As a physical endurance trial, the use of hot kola was supposed to form a man's character, to put iron into his soul.

**(u)laamba dibaghala** [ùlámà dibáyàlè]. "To cook a man". ■ To bewitch, to cast or put a spell on a man. Mi isalaamba babaghala! *I don't cast spells on men!*

**dibaghala na misopu** [dibáyàlè nà misópù]. "A male person with guts". ■ A courageous/brave man. Ω This expression is used with the verb «to be» (**uba**). Besides it may be used by a woman having an argument with her husband. Uba **dibaghala na misopu**. «*To have guts*». Ina dyaami **dibaghala na misopu**! Nyughumonisi dibaghala! *I'll show you my brother! I'll show a real man!*

**dibaghala ufutu mbaanda** [dibáyàlè ùfútù mbà:ndà]. "A man of whom the expert in this tradition of circumcision has spitted chilli (onto the penis during the circumcision test)". ■ A real man, a very courageous man. Nge la dibaghala ufutu mbaanda! Mwaana dibaghala aghu aghale! *You are a real man!*

## F

**FUUNDU/TSIFUUNDU** *n. cl. 9/10*

**(u)tuba o fuundu** [ùtúbà ò fú:ndù]. "To speak in secret assembly". ■ *To speak one's mind*. ⊗ This idiom is generally used in negative sentences. Duyatuba o fuundu! *Put the matter in the plain, speak frankly!* Mi isatuba o fuundu. Bak'utuba o fuundu a yooghu sa mi! *I use to look people straight in the eye and tell them what are my thoughts!*

## G

**GALU/TSIGALU** *n. (Loan) cl. 9/10*

**(u)tuula mu tsigalu** [ùtú:là mú tsígálù]. "To put one's child or someone in the braids". ■ To sacrifice one's child or someone to an evil genius. Batu bavatuba ti mwaana MAVHUUNGU wo atsifu mu maamba, a taayi atsimutuula mu tsigalu. *It is said that MAVOUNGOU's child who drowned in a boating accident was sacrificed by his father.* Anatuula mwaana mu tsigalu. *He killed his son and gave him to his gods.* ♦ Generally speaking, the Merye populations are driven by strong traditions, and despite the fact that most of them are now christianised, they still believe in spirits. These hidden forces, unknown to the layperson, would rule life and death, health and illness, prosperity and misfortune, etc. So in order **to be promoted**, certain persons do not hesitate to sacrifice their children or their close relations. From there is derived the saying "to put someone in the braids".

## I

**INDZABI** *n. cl. 7*

**(u)ghaka o Indzabi** [ùyákà ó òndzábì]. "To be bitten in Nzebiland". ■ To be or get caught by a nganga Kosi, to fall in his trap. Atsighaka o Indzabi. *The witch has been caught.* Bwaala buna bukaanda, atsighaka o Indzabi. *This village was strengthened medically, the witch has been caught.*

## K

**KUSU/TSIKUSU** *n. cl. 9/10*

**(u)ba na kusu** [úbà nà kúsù]. "To have a parrot (*Psittacus erithacus*)". ■ To master the art of oratory, public speaking. Mutu ghuna **ana kusu** tsyaandi tsi utuba. *This man is a smooth talker; he's got the gift of the gab.* MUSA AVHU **ana kusu**. MOUSSAVOU is one of the most eloquent of villagers. ♦ The parrot like the nightingale (**dutolu**) is considered to be a talkative, a garrulous bird which by the way makes him good company. In the

past, the parrots, in particular, were the victims of excessive hunting meant for the trade of animal species used as pets. Today, the grey parrot with a red tail (*Psittacus erithacus*), in particular, is part of the partially protected animals of Gabon. In most ethnolinguistic groups of the country, this species is regarded as sacred/holy. Among the Bayengui Koussou (clan of the parrot), it is forbidden to stamp on the excrement of the animal. Walking barefooted on parrot's droppings gives scabies to the Bayengui Koussou ("Bayeengi Kusu") as well as to their children ("baana ba Bayeengi"). It goes without saying that for this animal symbol or totem, it is strictly forbidden to the members of the Bayengui clan as well as to their progeny or offspring to eat the flesh of the parrot. The transgression of this food taboo is generally followed by serious consequences or repercussions.

## M

**MWAALA** *n. cl. 3*

**mwaala malibi** [mwá:là málìbì]. "High tide". ■

Said of a high tide-belly, high tidelike belly  
♦ Very complex and variable/changeable, tide depends on a certain number of factors, namely: the rotating movement of the earth around the sun, that of the moon round our planet, etc. From an empirical point of view, the ancients have already noticed that during the full moon (malunguna), in particular, the mass of water of the sea or the ocean is to a maximum (mwaala malibi). Scientifically, at very high tide (during the full moon or during the new moon), there are the attractions of the moon and the sun that are added. In the figurative meaning (sense obtained by figure of speech), mwaala malibi refers to a paunch or a potbelly, especially when talking about pregnant women. In the local French, the implicit comparison between very high tide, on the one hand and pregnancy on the other hand — as this is the case in Yilumbu — is replaced by another analogy

or metaphor: a pregnant woman with a paunch or potbelly is affectionately called ten wheels because her movements, bearing make people to think of her travel or travelling in terms of a lorry that is heavily laden.

## N

**NDZAANDA** *n. cl. 9*

**ndzaanda unangota mu miisu** [ndzã:ndà ùŋgótà mù mí:sù]. "Cobweb/spider's web went into my eyes". ■ To fall asleep. **Ndzaanda unangota mu miisu.** *I'm beginning to fall asleep.* **Ndzaanda unangota miisu.** *I can't keep awake, I can't keep my eyes open.*

## P

**PAGHASA/TSIPAGHASA** *n. cl. 9/10*

**(u)ba na dukotu du paghasa** [úbà ná dükótù dù páyàsà]. "To be (like) a wounded buffalo". ■ *To be cheating or double-dealing.* ♦ Under normal circumstances, the buffalo is a placid, a calm animal. However, when challenged, but above all when wounded, the buffalo has a terrible reputation of being two-faced. Hunters have learnt to their cost that a wounded buffalo can be very dangerous. In fact, he first runs away then he goes back the way in order to sound the charge. If the hunter manages to run away, often, the buffalo does not hesitate to follow him up to his door.

## S

**SAALU** *n. cl. 9*

**(u)tuula saalu** [ùtú:là sál:lù]. "To salt cases". ■ To exaggerate. Uyaghuulu mana, **atsituula saalu!** *The facts have been greatly exaggerated!* Nge una didodisi, mbaatsi asatuba nana, nge ika **saalu ukatuula!** *Don't exaggerate she didn't say that!*

## T

**TSESI/TSITSESI** *n. cl. 9/10*

**(u)baang'e tsesi** [ùbàngé tsé:si]. "To be clever as a gazelle". ■ To be very clever; to be a fox (someone who is clever and good at deceiving people). **Abaang'e tsesi**. Ukebe! *He is very clever. Be careful!* Mutu ghuna **abaang'e tsesi**. *This man is a cunning old fox.* ♦ In tales, the gazelle, shrewd (cunning), is always on her guard, which warrant her to triumph over the underhand tricks of predators (in particular, the leopard nicknamed "ya Ngo").

## U

**UTALA** *n. cl. 15*

**(u)tala vhotsi** [ùtálà βó tsì]. "To look down, on the ground". ■ To be ashamed. Tsoni tsinamusiimba **anatala vhotsi**. *He is so ashamed of what he did and can no longer look people straight in the eye.*

**USEVHA** *n. cl. 15*

**(u)sevhá bakaata** [ùséβà bákà:tà]. "To laugh at elders". ■ To be impudent. NGOMA sana butu, **avasevha bakaata**! *NGOMA is such an impudent child!* **Avasevha bakaata**, avanata musabu, isumaaghu. *Instead of eagerly serving his elders, he's showing disrespectful behaviour.*

## V

**VUSUU** *onomat.*

**(u)si vusuu na ubura** [úsi vúsú: nà úbùrà]. "To do + onomatopoeia expressing completeness". ■ To be fertile. MANOOMBA **asi vusuu na ubura**. *MANOMBA has many children.* Vhana **asi vusuu na ubura**, kadi muvhu mwaana, kadi muvhu mwaana. *She is so fertile that each year she is giving birth to a child.*

## W

**WALI** *n. cl. 9*

**(u)ba mu wali** [úbà mù wáli]. "To be in the *Garcinia klaineana* tree". ■ To be drunk. **Bali mu wali**. *They are drunk.* **Wali** yikamumagha o muru. *He is under the influence of intoxicating drink.* ♦ People make palm wine by leaving the bark of the *Garcinia klaineana* tree to ferment inside palm juice.

## Z

**ZORU** *n. cl. 9*

**(u)ba na buvantaara bu Zoru** [úbà ná búvántà:rà bú zòrù]. "To have the boastfulness of Zorro". ■ To be too full of oneself. ⊗ This idiom is generally used as a mocking expression. Ana **buvantaara bu Zoru**. *He likes praising himself, he is too full of himself.* Botugha vhana na **buvantaara bu Zoru**! *It annoys me that you enjoy boasting about your achievements!* ♦ According to the legend, Zorro (male fox, cunning fellow, in Spanish) was the defender of the poor and the oppressed. In the state of California, he used to steal from the rich and give to the poor. The different movies that were made out of that legend were very successful in Gabon and in Libreville in particular. In this town, Zorro, a notorious policeman, was named after the famous Zorro of the movies not because of his qualities of righter of wrongs but because he used to terrify and brutalise people who sell goods illegally on the streets of Libreville's market places.

## CHAPTER 12

# LANGUAGE STANDARDISATION AND OTHER LANGUAGE AND CONTENT RESOURCES: SABS TC 37 - A MIRROR COMMITTEE OF ISO/TC 37

Mariëtta Alberts

*Research Unit for Languages and Literature in the SA Context, North-West University,  
Potchefstroom, South Africa  
malberts@lantic.net*

### 1. INTRODUCTION

Standardisation is defined as an “*activity of establishing, with regard to actual or potential problems, provisions for common and repeated use, aimed at the achievement of the optimum degree of order in a given context*” (ISO/IEC 1996) (cf. ISO/TC 37 2004).

Standardisation comprises a whole world of activities in various spheres of life which helps creating benefits and sustainability for everyone in the global society. Standards make things easier for all, e.g. electric plugs need to adhere to specific standards to be used locally and/or internationally – adaptors that comply with specific national and international standards are sometimes needed as a result of plugs that do not adhere to international standards (Alberts 2008a:2).

The benefits of standardisation:

- simplification
- economical (cost saving)
- interoperability of systems and components
- health and safety
- consumer protection
- the elimination of trade barriers
- better communication (cf. Alberts 2008a:2)

This chapter focuses on the role and value of standardisation in language-related environments and special emphasis is placed on terminology and other language and content resources.

Terminology plays a pivotal role wherever and whenever specialised knowledge (information) is generated or prepared (e.g. in research and development),

recorded and processed (e.g. in databanks), transferred (via teaching and training), used (e.g. in specialised texts), implemented (e.g. in technology and knowledge transfer), or translated and interpreted. Terminology facilitates language development, the creation of knowledge and the dissemination of information. Terminology is indispensable in a multilingual global information and knowledge environment where this society depends on reliable digital content. Information on science, technology, economy and all aspects of human endeavours is disseminated by means of terminology. In other words: No knowledge on any subject field or domain can be disseminated without terminology (cf. Alberts 2006a & 2008a:1). The TermNet slogan says it all: *No knowledge without terminology*.

The terminology practice, however, depends on standards – albeit standard terminological principles, practices and procedures or the terminological data itself. Standards, however, play an important role in the capturing and the dissemination of all linguistic data – albeit for lexicography (general words), terminography (multilingual polythematic terms), translation work (translation memories) or for the various fields of human language technologies (HLT). These standards relate to principles, methods, applications and the data itself. While lexicography deals with the documentation of general words and has a descriptive approach, terminography deals with subject-related terms and follows a prescriptive approach since terms should be standardised to ensure exact communication. No dictionary, whether general or for special purposes, can be compiled without standardised grammar, spelling and orthography rules. The languages chosen as standard varieties need to adjust since no language is static, languages change and modernise and therefore these standard languages need to adapt to change. This chapter focuses on the standardisation of language and issues relating to the work of the international standards committee, ISO/TC 37, dealing with terminology and other language and content resources as well as the South African mirror committee, SABS TC 37.

## **2. LANGUAGE DEVELOPMENT AND THE STANDARDISATION, MODERNISATION AND HARMONISATION OF LANGUAGES**

Chumbow and Tamanji (1998:53) use the term ‘language development’ to refer to all language engineering activities undertaken to bring a language from its natural state (the oral form in a pre-literacy state) to a codified and standardised form with some amount of literature and literacy activities as well as vibrant communication activities in terms of the conveyance of knowledge in various aspects of modernism, science and technology.

The majority of languages are composite varieties characterised by multiple selections, i.e. the complex recombination of features from various dialects and

language varieties. Standardised languages have multiple ancestors and their history is shaped by various types of language content.

Standard languages are usually associated with prestige and cut across regional differences, providing a unified means of communication. Standard languages could come into existence spontaneously through historical and social circumstances (e.g. the regional form of English of the London area was used for higher functions (e.g. law) and printing and became Standard English. Standardisation could therefore be a **natural development** of a language into a standard language within a speech community, i.e. the development of Afrikaans from “kitchen language” into a language with the status of “standard language”, fit for higher functions. A standard language could, however, also be created intentionally with an attempt by a community to **impose one dialect** (e.g. Sepedi) as a standard (Sesotho sa Leboa). Standardisation could also be a **direct and deliberate intervention** by a society to create a standard language where before there were just dialects (non-standard varieties) as is the case with Kiswahili. The standard language needs to be backed up by the linguistic community and the educational system (Alberts 2006a).

A standard language is a language empowered by its linguistic community to fulfil the functions which would make it a standard language (cf. Kotzé 2009:1). The process of standardisation traditionally follows a particular route, i.e. via norm selection, codification of form, elaboration of function and acceptance by the community (cf. Haugen 1974 in Kotzé 2010). Although it is assumed that the patterns of usage of an idealised group of speakers normally serve as basis for the standard form, the standard language is at most a stylised register for specific purposes (Kotzé 2010).

Standard languages are often regarded as the only variety of importance – being ‘pure’, ‘proper’ and ‘correct’. A standard language could be used for social manipulation and could be seen as an elitist tool and instrument of alienation and emotional insecurity and language-internal conflict. It could also be seen as an exclusive language of modernisation and the only vehicle for development (cf. Webb 2005: 39-40, Alberts 2006a).

Two broad types of processes can be distinguished in the standardisation of languages:

- Social processes – these concern the modification to the status of the particular variety adopted as the standard in a given speech community.
- Linguistic processes – these concern developments within the corpus of the language itself, i.e.
  - elaboration of function (lexicon)

- codification (written form, grammar, syntax, etc.) (Alberts 2006a).

The standardisation of a language is a language planning concern and in general a sociolinguistic and political issue. It is concerned with the linguistic form (corpus planning) and the social and communicative functions of language (status planning) (cf. Alberts 2006a, Kotzé 2009:1).

Standardisation of languages involves the progressive elimination of alternative norms through the selection of one norm, which is then superimposed on the rest. This selection process could be described as a language external process. Selection could, however, also be a language internal process where specific language elements need to be standardised according to certain formal requirements. The language external and internal processes are inter-dependant but can also function separately. In the process of standardisation there is often a coalescence of the internal and external history of the language (cf. Kotzé 2009:1).

Standardisation also involves:

- geographical distribution and regional variations
- recognition of the standard variety
- influences of other languages which might be enriching or detrimental
- acceptance by the speakers
- status, value and usefulness as a fully-fledged language, and
- development (e.g. terminology development) and promotion of the language (cf. Alberts 2006a).

Standardisation is not a categorical variable, but it exists in many gradations ranging from full standardisation of a language (showing uniformity of usage, diversity of functions and full ideological rationalisation) to various forms of partial or incipient standardisation (i.e. pidgins and creoles) (cf. Alberts 2006a). There are also processes such as de-standardisation and re-standardisation influencing the status of a chosen standard variety (cf. Kotzé 2009:1).

Standardisation is necessary –

- to facilitate communication
- to make possible the establishment of an agreed spelling and orthography
- to provide a norm for the written form of the language (e.g. for literature, education, translation, journalism, etc.), and
- to provide a uniform form for dictionaries (general lexicography) and terminology development (terminography) (cf. Alberts 2006a).

The role of standard languages is

- to facilitate effective communication across dialectal differences
- to be used in all official documents to facilitate communication between government and citizens
- to be used in high-function formal contexts (parliament, legislation, court, business)
- to be used as languages of instruction, in education, science, economy, technology, etc.
- to have a symbolic function by generally representing the linguistic identity of the members of a particular language community (Alberts 2006a).

A language is regarded as standardised when it possesses a grammar, spelling and orthography and dictionaries. The grammar, spelling and orthography should be fairly settled. The language should be widely understood and the vocabulary should encompass every possible function (e.g. legal, commercial, academic, scientific, and technological). A standard language should be teachable on primary, secondary and tertiary level and a variety of general and special purpose dictionaries should be available (Alberts 2006a).

A standard language is codified and uniform. Languages, however, change and develop. Language standards have to be updated and revised frequently because of language change and the fact that languages are modernised. Standardisation, therefore, is an ongoing process and linguistic communities and governing bodies (i.e. National Language Bodies) therefore frequently need to determine the status of standardisation of the chosen standard varieties, e.g. the various official South African languages and other indigenous languages.

One of the linguistic aspects that needs frequent attention is the revision of spelling and orthography rules of standard languages. Language-related professions such as lexicography, terminography and translation work adhere to the spelling and orthography rules of the standard variety of the language. New terms that are created in a language, need a solid linguistic basis with proper word-forming principles as norm/standard.

Terminology is the medium through which knowledge and information is disseminated. The use of standardised terminology results in effective communication. The multilingual global environment with a plethora of languages gives rise to challenges regarding communication and even training. The harmonisation of conceptual structures, conceptual systems and term equivalents could assist with term creation problems (terminology development) and the



dissemination of polythematic information and knowledge (cf. Alberts 2006b:185-188).

Standardisation and harmonisation both involve a direct and deliberate intervention by the linguistic society and governing bodies to create a neutral standardised variety. Harmonisation should be a natural process and should not be enforced on any language or language community. The terminography process could, however, benefit from the principles and practice of harmonisation – especially in a multilingual environment.

Not only should terminologies be harmonised and standardised to ensure exact communication over a broad spectrum of languages and domains, but care should be taken that problems associated with cultural sensitive concepts should be addressed. Socioterminology has an important role to play in the global society where information transfer occurs from one language and one culture to another. What is regarded as offensive or taboo in one language may be a term commonly used in another language and culture. This problem could be solved by using transliterations or euphemisms (cf. Alberts 2006b:188-190).

### 3. TERMINOLOGY STANDARDISATION

ISO 1087-1:2000 (ISO/TC 37 2000) defines **terminology** as “a set of designations belonging to one special language” and **designation** as the “representation of a concept by a sign which denotes it.” There should be a one to one relationship between concept and term to ensure exact communication – without standardised terms confusion could result.

A **terminology standard**, according to ISO/IEC Guide 2 (1996), is defined as a “standard that is concerned with terms, usually accompanied by their definitions, and sometimes by explanatory notes, illustrations, examples, etc.” (cf. ISO/TC 37 2004). Recently, terminology standardisation has been subdivided into the following two distinct categories:

- standardisation of terminologies, and
- standardisation of terminological principles and methods (cf. Alberts 2008a:2).

The beginnings of terminology standardisation are closely linked to the standardisation efforts of the International Electrotechnical Commission (IEC), founded in 1906, the International Federation of Standardizing Associations (ISA), founded in 1926, and the International Organization for Standardization (ISO), founded in 1946. The IEC considered its foremost task to be the standardisation of the terminology of electrotechnology for the sake of quality of its subject standards. Eugen Wüster's book of 1931 “*Internationale Sprachnormung in der Technik*” (International standardisation of technical language) triggered the establishment in

1936 of the Technical Committee ISO/TC 37 “Terminology” for the sake of formulating general principles and rules for terminology standardisation (cf. ISO/TC 37 2004, Alberts 2008a:2).

ISO standardised terminology comprises all terms and definitions occurring in ISO standards. As such standards are an important element of teaching and training of subject-related topics, the initiation phase into any specialised field goes through the learning of the main concepts (and the terms and definitions describing them) (cf. Alberts 2008a:7).

#### **4. STANDARDISED COMPUTER APPLICATIONS**

The global accessibility of data makes it clear that computer accessibility and other electronic devices have a major influence on the working environment, education, private life and society as a whole. There is an increased level of global exchanges, standardisation and interdisciplinary approaches. Lexicographers and terminographers likewise explore what effects these innovations have on their particular trade. They need to consult lexicographical and terminological data banks and need to use computerised material for the compilation of dictionaries and related products and use computerised means to publish and disseminate linguistic information.

Although computerisation brought huge advantages for general lexicography and terminography there are a variety of problems which resulted as a consequence of computer-assisted work. Some areas of general lexicography and terminography which need to be addressed are (cf. Alberts 2010d):

- protected material, legislation, and copyright issues (cf. Alberts & Jooste 1998:122 -139, Mfafa 2003:118-122);
- selection – which items to include (and providing criteria for inclusion such as frequency); establishment of a headword list; offering headword list which may readily be reduced; criteria for headword status (compounds, phrases, idioms, etc.);
- definition (monolingual dictionaries) or translation (bilingual and multilingual dictionaries) and providing access to reliable information; control of defining vocabulary; avoidance of circular definitions, etc.;
- storage, manipulation and retrieval of phonological, syntactical, semantic, morphological and pragmatic information;
- consistency of style labels (register, geographical, taboos, etc.);
- validating of cross-references; cross-checking in bilingual and multilingual dictionaries of items of source and target languages;

- consistency of individuals' work in a large team of compilers;
- provision of corpus of written and spoken language;
- ready reference to instances of use in corpus, by means of concordancing;
- flexibility of databank and consequent enhanced potential (cf. also Atkins 1982:261-262);
- production of satellite products from a single databank;
- availability and compatibility of different kinds of software (cf. Alberts 2002:96-97).

The South African Government has approved the development of a human language technologies (HLT) virtual network which is indicative of the importance of applications of terminology and other language resources in language engineering and content management. All lexicographical and terminological endeavours will be part of the HLT virtual network. The multilingual linguistic data will be available to end-users such as subject specialists, students, language practitioners, and the general public (cf. DAC 2002-2004, Alberts 2010b:618, Alberts 2010c).

The application of HLT makes it possible for human and computer interaction by using human language (text and speech). These technologies range from high-level parsing and machine translation to voice-operated educational or commercial systems that can be used by illiterate people, and from applications in education and training to public service (e-governance) and e-commerce applications. These enabling HLT applications provide human-computer interfaces and linguistic assistance such as spelling and grammar checking. They are, however, dependent on the availability of large collections of electronic linguistic data (e.g. word and term banks). HLT initiatives need data from a variety of sources, which are compatible, standardised and reusable. These resources should conform to international HLT and ISO/TC 37 standards and meet a vast variety of needs in the multilingual society (cf. Alberts 2010c & 2010d).

## **5. ISO TC/37: TERMINOLOGY AND OTHER LANGUAGE AND CONTENT RESOURCES**

Standards play an important role since (inter)national standards can even be incorporated into legislation by referring to them. Compulsory specifications can be redrafted as legislation. All standards are voluntary unless incorporated into legislation. As standards are second to written law, they affect all aspects of human life. (cf. Alberts 2008a:7).

The Technical Committee (TC) ISO/TC 37 deals with the standardisation of principles, methods and applications relating to terminology and other language and content resources in the context of multilingual communication and cultural diversity (cf. ISO/TC 37 2004, Alberts 2008a:3).

ISO/TC 37's **objective** is to prepare standards that specify principles and methods for the preparation and management of terminology, language and other content resources (at the level of concepts) within the framework of standardisation and related activities. Its technical work results in International Standards (and Technical Reports), which cover terminological principles and methods and also various aspects of computer-assisted terminography. However, ISO/TC 37 is not responsible for the co-ordination of the terminology standardising activities of other ISO/TCs (cf. ISO/TC 37 2004, Alberts 2008a:3).

ISO/TC 37's **mission** is to provide standards and guidelines to standardisation experts, language professionals in all institutions and organisations that create and handle terminology, language and other content resources (including ISO itself, other international organisations, national standards bodies, national government services, companies, non-governmental organisations, etc.) in order to enable them to prepare high-quality language resources and tools for a whole variety of applications in professional and scholarly information and communication, education, industry, trade, etc. (cf. ISO/TC 37 2004, Alberts 2008a:3).

Part of ISO/TC 37's **vision** is that worldwide use of ISO/TC 37 standards will help to enhance the overall quality of terminologies and other language and content resources in all subject fields, to improve information management within various industrial, technical and scientific environments, to reduce their costs, and to increase efficiency in technical standardisation and professional communication (cf. ISO/TC 37 2004, Alberts 2008a:4).

Currently ISO/TC 37 has 27 participating member countries (**P** membership) and 35 Observing countries (**O** membership). It is in liaison with 34 international organisations and in internal liaison with 26 technical bodies of ISO, IEC and the European Committee for Standardisation ((Comité Européen de Normalisation (CEN)). Sixteen standards have been published and 28 working items are in progress (cf. ISO/TC 37 2010).

The **secretariat** of ISO/TC 37 was until the end of 2008 operated by Infoterm, the International Information Centre for Terminology, on behalf of the Austrian Standards Institute (ON). The Secretariat of ISO/TC 37 was then transferred from Infoterm (Austria) to its Twinning Secretariat, the China National Institute of Standardisation (CNIS) (China). ISO/TC 37 operates in two **official languages** namely English and French (cf. ISO/TC 37 2004, Alberts 2008a:4).

ISO/TC 37 functions according to subcommittees and working groups. It is overseen by an Advisory Group (TC 37/AG) and currently has four subcommittees (SCs) (cf. ISO/TC 37 2004, Alberts 2008a:4).

### 5.1. ISO/TC 37/SC 1: Principles and Methods

The **scope** of ISO/TC 37/SC 1 is the standardisation of principles and methods relating to terminology, language resources, terminology policies and to knowledge organisation in the mono- and multilingual context of the information society (cf. ISO/TC 37 2004, Alberts 2008a:4, 2010a). The **objective** of ISO/TC 37/SC1 is to prepare standards that lay down the basic principles for preparing, updating and harmonising terminologies and other language and content resources on the one hand, and to standardise principles and methods related to terminology policies and knowledge organisation in the multilingual information society (cf. ISO/TC 37 2004, Alberts 2008a:4; 2010a). ISO/TC 37/SC1's **mission** is to provide standardisation experts of national and international standards bodies and language professionals in international organisations, national government services, companies, non-governmental organisations, etc. with relevant standards and guidelines to assist them in creating high-quality terminologies and other language and content resources, and in formulating terminology policies and implementing knowledge organisation (cf. ISO/TC 37 2004, Alberts 2008a:4, 2010a).

### 5.2. ISO/TC 37/SC 2: Terminographical and Lexicographical Working Methods

The **scope** of ISO/TC 37/SC2 is the standardisation of terminographical and lexicographical working methods (e.g. translation-oriented terminography), procedures, coding systems, workflows, and cultural diversity management (e.g. standardisation of two and three letter abbreviations of language and country names), and also related certification schemes (cf. ISO/TC 37 2004, Alberts 2008a:5). The **objective** of ISO/TC 37/SC2 is to prepare practice-oriented standards for terminology work, terminography, lexicography, and reference coding. ISO/TC 37/SC2 will pursue this objective by: identifying and targeting the client audience, making the standards available on the market, and identifying and meeting client needs (cf. ISO/TC 37 2004, Alberts 2008a:5). The **mission** of ISO/TC 37/SC2 is to provide advice concerning terminological and lexicographical working methods, procedures, coding systems, workflows, and cultural diversity management, and also related certification schemes through the publication of standards and the use of the internet in order to meet the needs of the client audience (cf. ISO/TC 37 2004, Alberts 2008a:5).

### 5.3. ISO/TC 37/SC 3: Systems to Manage Terminology, Knowledge and Content

The **scope** of ISO/TC 37/SC3 is the standardisation of specifications and modelling principles for systems to manage terminology, knowledge and content with respect to semantic interoperability (cf. ISO/TC 37 2004, Alberts 2008a:5). The **objective** of ISO/TC 37/SC3 is to develop standards for the sake of semantic interoperability comprising specifications of terminology, language and content management, which cover data modelling, mark-up, data exchange, and evaluation of terminology management and knowledge ordering tools (cf. ISO/TC 37 2004, Alberts 2008a:5). The **target groups** of SC3 are providers and users of terminology, language resource, content and knowledge management, including software companies active in this field. The scientific community catering to those services belongs to these target groups as well as educational institutions (cf. ISO/TC 37 2004, Alberts 2008a:5).

### 5.4. ISO/TC 37/SC 4: Language Resource Management

The **scope** of ISO/TC 37/SC4 is the standardisation of specifications for computer-assisted language resource management (cf. ISO/TC 37 2004, Alberts 2008a:5). The **objective** of ISO/TC 37/SC4 is to prepare standards by specifying principles and methods for creating, coding, processing and managing language resources, such as written corpora, lexical corpora, speech corpora, dictionary compiling and classification schemes. These standards will also cover the information produced by natural language processing components in these various domains. Standards produced by ISO/TC 37/SC4 particularly address the needs of industry and international trade as well as the global economy regarding multilingual information retrieval, cross-cultural technical communication and information management (cf. ISO/TC 37 2004, Alberts 2008a:5). The **goal** of ISO/TC 37/SC4 is to ensure that new developments in language engineering, knowledge management and information engineering satisfy the norms of international standardisation. This can be achieved through the development of standards and related documents to maximise the applicability of language resources of different kinds, relating them to applications, and enhancing the application of recognised methods and tools in language resources (cf. ISO/TC 37 2004, Alberts 2008a:6).

## 6. SABS TC 37: TERMINOLOGY AND OTHER LANGUAGE AND CONTENT RESOURCES

South Africa participated in international ISO/TC 37 activities for many years but these activities were not sustained. A national ISO/TC 37 committee, StanSA TC 37, in the field of Terminology and other language resources was established on 19

August 2002 (cf. Alberts 2008a:6). Prof Justus Roux was elected as the first chairperson of this committee. The name of the committee later changed to SABS TC 37 and this committee functions as the local standardisation committee regarding terminology and other language and content resources. SABS TC 37 is a mirror committee of the international ISO/TC 37. The current chairperson is Dr Mariëtta Alberts. The administration of SABS TC 37 is dealt with by the South African Bureau of Standards (SABS), Groenkloof, Pretoria (cf. Alberts 2008a:6 & 2008b).

SABS TC 37 initially held an O (Observer) membership status at international level but in 2004 obtained a P (Participant) membership status (cf. SABS 2011). Members of SABS TC 37 represent South Africa at the annual international ISO/TC 37 meetings. The SABS TC 37 team has a very valuable role to play in deliberating at international level when standards dealing with terminology, lexicography, translation work and other language and content resources are under revision. They also gain valuable experience from the international scene, which they are then able to share with their South African colleagues (cf. Alberts 2008a:9). SABS TC 37 formed national liaisons with SABS TC46: Information and Documentation and SABS TC170: Knowledge Management (SABS 2011).

There are principles regarding standards that are of the utmost importance when the development of standards is considered. National standards are developed by consensus through technical committees, representing the needs of local stakeholders, bearing in mind the type of discipline, modern technology, programme management, and cost effectiveness. Committees are comprised of a balanced representation of user groups, interest groups as well as government. New standards have the advantage that they could focus on addressing the needs of the users of the standards, but the process is very time consuming and costly. Adopting regional and international standards has the advantage of having standards that were well researched by international experts and are moderately priced (cf. Alberts 2008a:7 & 2008b). All published standards are subject to routine maintenance by circulating questionnaires to the relevant committees every five years to reaffirm the validity of the standards, to amend, to revive or to withdraw them (cf. Alberts 2008a:7). The business environment of SABS TC 37 includes the following aspects:

- Globalisation
- National language policy framework
- Regionalisation and co-operation possibilities within the Southern African Development Community (SADC)
- Localisation and centralisation
- Promotion of multilingualism

- Human language technology (HLT)
- Promotion of the values of standardised terminology principles within business, industry, science and technology (cf. SABS 2011:1).

The scope of SABS TC 37 is the standardisation of principles, methods and applications relating to terminology and other language and content resources (SABS 2011). The committee organises annual workshops to address one or more aspects of its scope to ensure national participation in its activities. The committee also compiled a technical brochure (cf. Alberts 2008a) and a flyer to disseminate information on SABS TC 37.

The national committee, as mirror of the international committee, concentrates on four subcommittees (SCs) (cf. Alberts 2008a:6 & 2008b) and the work relating to their working groups (WGs), namely:

- Subcommittee (SC) 1 “Principles and methods”
  - Scope: Standardisation of principles, methods, and applications relating to terminology and other language and content resources.
- Subcommittee (SC) 2 “Terminology and lexicography”
  - Scope: Standardisation of the application of principles and methods in terminology work, terminography and lexicography.
- Subcommittee (SC) 3 “Computer applications for terminology”
  - Scope: Standardisation of models for information processing and of related coding systems applicable to terminology work and terminography.
- Subcommittee (SC) 4 “Language resource management”
  - Scope: Standardisation of specifications for computer-assisted language resource management.

The SABS Technical Committees (TCs) and Subcommittees (SCs) take part in the work of relevant ISO committees and act as mirror committees. No projects on local standards are undertaken without considering the ISO work in progress or published standards. It remains the decision of the local committee (e.g. SABS TC 37) to adopt an international standard as a national standard (SANS) or to write a standard from scratch if the international standard is not suitable or relevant (cf. Alberts 2008a:7 & 2008b).

The attendance of international meetings provides great opportunities to meet up with colleagues working in the same domain and sharing ideas. The limited participation of South African language practitioners in international activities in the field of ISO/TC 37 is logistically and geographically related. Their influence is, however, clearly recognised at meetings of this nature. Following various



discussions at the recent 2010 ISO/TC 37 plenary sessions in Dublin, Ireland, it was pointed out by international colleagues that the South African Committee system (i.e. SABS TC 37) and its annual activities (research, meetings and workshops) were rather unique compared to what is taking place in other countries. "The role of South African influence in a direct as well as indirect manner in international standardisation processes should not be under-estimated" (SABS 2010:10). The SABS TC 37 committee envisages liaison and collaboration across regional borders (e.g. through SADC and the African Academy of Languages (ACALAN)) and will host the ISO/TC 37 plenary sessions in 2013 in South Africa.

## 7. CONCLUSION

Standardisation is an ongoing process since languages are dynamic, they modernise and change. Language activists, language communities and stakeholders should keep track of changes in respective languages. The SABS TC 37 committee aims to involve all language practitioners from various institutions into language related standardisation processes. The national committee mirrors the work of the international ISO/TC 37 and through regular liaison, communication and feedback global activities regarding terminology and other language and content resources could be shared.

## REFERENCES

- Alberts, M. 2002. E-Terminology. *Lexikos* 12:90-104.
- Alberts, M. 2006a. *Standardisation of languages*. Paper presented at the 11<sup>th</sup> International Afrilex Conference, organised by the Tshivenda National Lexicography Unit, University of Venda for Science and Technology, Thohoyandou, RSA, 5-7 July 2006.
- Alberts, M. 2006b. Standardisation, Modernization and Harmonization – A Sociolinguistic perspective. *Proceedings TSTT'2006 International Conference, 26 – 27 August 2006 on Terminology, Standardisation and Technology Transfer*, edited by Yuli Wang. Beijing, China: Encyclopedia of China Publishing House. 182-192.
- Alberts, M. 2008a. *Technical document. Terminology & Standards. Technical Committee (TC 37)*. Technical document. SABS. Groenkloof, Pretoria.
- Alberts, M. 2008b. *SABS TC 37 – an overview*. Paper presented at the SABS TC 37 workshop, 24 October 2008, SABS, Groenkloof, Pretoria.
- Alberts, M. 2010a. *SABS TC 37 SC1 Principles and Methods*. Paper presented at the SABS TC 37 workshop, 5 February 2010, SABS, Groenkloof, Pretoria.

- Alberts, M. 2010b. National Language and Terminology Policies – A South African Perspective. *Lexikos* 20:599-620.
- Alberts, M. 2010c. *Module TPL 754 Electronic terminology management*. University of the Free State, Bloemfontein.
- Alberts, M. 2010d. *Module TPL 774 Practising lexicography with the aid of corpora*. University of the Free State, Bloemfontein.
- Alberts, M. & M. Jooste. 1998. Lexicography, Terminography and Copyright. *Lexikos* 8:122-139.
- Atkins, B.T. Sue 1982. Remarks made at the Symposium. *Lexicography in the electronic age. Proceedings of a Symposium held in Luxembourg, 7 – 9 July, 1981*, edited by Goetschalckx, J. and Rolling, L. Amsterdam: North-Holland Publishing Company.261-263.
- Chumbow, B.S., Tamanji, P.N. 1998. Linguistic identity across the borders of the Cameroon triangle. *Between distinction & Extinction. The Harmonisation & Standardisation of African Languages*, edited by Prah, K.K. Johannesburg: Witwatersrand University Press.53-74.
- Department of Arts and Culture. 2002. *Report of the Advisory Panel on the Development of Human Language Technologies Progress Report. 8 November 2001 - 26 August 2002*. National Language Service, Department of Arts and Culture.
- Department of Arts and Culture. 2003. *Progress report on Human Language Technologies project. Human Language Technologies. 4/4/B*. National Language Service. Department of Arts and Culture.
- Department of Arts and Culture. 2003. *Report Workshop on Human Language Technologies (HLT) held on 28 February 2003 at the Holiday Inn, Beatrix Street, Arcadia*. National Language Service. Department of Arts and Culture.
- Department of Arts and Culture. 2004. *Draft Human Language Technologies (HLTs) Policy Document*. National Language Service, Department of Arts and Culture.
- ISO/TC 37. n.d. *ISO/TC 37 flyer*. Vienna, Austria: Austrian standards Institute in co-operation with the Standardisation Administration of China (SAC). <http://www.iso.org>. Accessed: 01-04-2006.
- ISO/TC 37. 2000. *ISO 1087-1:2000. Terminology Work – Vocabulary – Part 1: Theory and Application*. International Organization for Standardization.
- ISO/TC 37. 2004. *50 Years ISO/TC 37 "Terminology and other language and content resources"*. ISO/TC 37N 499; ISO/TC AG N 117. Geneva: ISO.

- ISO/TC 37. 2010. *Updated draft of ISO TC 37 Business Plan 2010-2011*. Report (2010), No 1852. [http://isotc.iso.org/livelink/livelink/open/tc37\\_bp](http://isotc.iso.org/livelink/livelink/open/tc37_bp) Accessed: 01-03-2011.
- Kotzé, E.F. 2009. *Hoe standaard kan 'n taal wees? Perspektiewe oor die teenstrydighede van Afrikaans*. Paper presented at the Roots Conference, *Nuwe roetes na nuwe wêreld: Spreek – Thetha – Talk: 'n Suid-Afrikaans-Nederlandse dialoog oor die dinamika van taal, kultuur en erfenis*. 22 – 23 September 2009. University of the Western Cape, Bellville, Cape Town.
- Kotzé, E.F. 2010. Destandaardisasie en herstandaardisasie – gelyklopende prosesse in die Nuwe Suid-Afrika? *Standaardtalen in beweging*, edited by Van der Wal, Marijke and Francken, Eep. Amsterdam, Stichting Neerlandistiek VU & Münster: Nodus.
- Mfafa, X.T. 2003. Terminology Coordination and Copyright Issues. *TAMA 2003 South Africa Terminology in Advanced Management Applications. 6<sup>th</sup> International TAMA Conference: Conference Proceedings*, edited by De Schryver, Gilles-Maurice. Pretoria: (SF)<sup>2</sup> Press.118-122.
- SABS. 2010. *ISO/TC 37 Meetings and Plenaries held in Dublin, Ireland, 16-20 August 2010*. SABS TC 37 Report (2010), No 1852. SABS. Groenkloof, Pretoria.
- SABS. 2011. *Strategic Policy Statement. TC 37: Terminology and other language and content resources. As amended after the meeting on 13 February 2009*. [http://www.sabs.co.za/Business\\_Units/Standards\\_SA/Controls/SPS\\_HTML/37.html](http://www.sabs.co.za/Business_Units/Standards_SA/Controls/SPS_HTML/37.html) Accessed: 01-03-2011.
- The Advisory Panel, Minister of Arts, Culture, Science and Technology. 2000. *Language Policy and Plan for South Africa. Final draft. 29 February 2000*. Unpublished DACST document.
- TermNet - International Network for Terminology. n.d. <http://www.termnet.org> Accessed: 01-03-2011.
- Webb, V.N. 2005. The Role of Language Standardisation in the Effective Functioning of Communities in Public Life in South Africa. *The Standardisation of African Languages. Report of the Workshop held at the University of Pretoria, 30 June–1 July 2005*, edited by Webb, V., A. Deumert and B. Lepota. Pretoria: PanSALB.35-42.

## CHAPTER 13

### ENRICHING A DICTIONARY DATABASE WITH MULTI WORD EXPRESSIONS

Thapelo Otlogetswe

*Department of English, University of Botswana, Gaborone, Botswana*  
otlogets@mopipi.ub.bw

#### 1. INTRODUCTION

A dictionary's primary role is the description of linguistic units, that is, a language's lexicon (Svensén 2009). Such knowledge representation includes a word's meaning, its conventional spelling (which may include regional spelling variations such as colour/color, behaviour/behavior), pronunciation (which may also include pronunciation variations as in the various possible pronunciations of the words 'either' and 'potato'), its syntagmatic categories (the behaviour of a word in combination with other words both grammatically and lexically) and the various inflections that a headword may take (Grefenstette 2009:307; Béjoint 2010). The headword's lexical relations may also be captured not only to differentiate it from other words but also to strengthen a user's vocabulary knowledge. Such lexical relations may include synonymy, antonymy, meronymy and others.

However, even with such detailed lexical information, there is still much more to a headword that may be added in a dictionary to enhance a dictionary's functionality. Such additional information in a dictionary has been termed "usage notes" in the *New Oxford Dictionary of English* (Pearsall 1998) and *Longman Dictionary of Contemporary English* (LDCE) (Summers 1995). Usage notes take the form of information of how an entry is used in the language. They advise users on typical mistakes that they are likely to make, e.g. confusing 'adequate', 'sufficient' and 'enough', 'good enough', 'satisfactory' and '(will) do'. Some additional information, such as in the LDCE, includes frequency information of a headword in specific domains, such as in the spoken and written domains. A dictionary may also be enriched with the addition of idioms, multiword units and collocates. This enriching of the dictionary is necessary since a dictionary is often seen as a list of simple words.

However, in linguistic literature, the term *word* is a contested term. Some of the definitions of wordhood, while useful for theoretical linguistics, are not helpful for computational word counts and lexical computing in general. Finch (2000:132) defines a word as "a unit of expression which native speakers intuitively recognise in both spoken and written language" and adds that "*there is a certain indeterminacy*

about the definition of a word". In most computational processes, a word is treated as a "minimal free form, the smallest unit that can exist on its own" (Dash and Chaudhuri 2000:189) and "delimited by a space ... on each side" (Leech *et al.* 1982:27). This approach is helpful if one is studying forms delineated by spaces only. However, limitations of perceiving a word as a string of characters delimited by spaces is well known (Cowie 1998; Otlogetswe 2009) since there are larger linguistic units which have spaces within them which function as lexical units. These larger units are the subject of this paper. Moon (1998) calls such units fixed expressions and idioms. In other literature they are called multi-word units (MWUs) (Schone and Jurafsky 2001) or multi-word expressions or (MWEs) (Sharroff 2004; Oflazer and Çetinoğlu 2004; Villavicencio *et al.* 2004; Fazly and Stevenson 2007). Bannard (2007:1) defines a multi-word unit as:

...any word combination (adjacent or otherwise) that has some feature (syntactic, semantic or purely statistical) that cannot be predicted on the basis of its component words and/or the combinatorial processes of the language. Such units need to be included in any language description that hopes to account for actual usage.

Sag *et al.* (2002:2) characterise MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)". Jackendoff (1997:156) has estimated that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words. If his estimate is accurate, then MWEs deserve focus and will significantly enhance dictionary entries.

MWEs therefore include idioms, phrasal verbs, proverbs, compound words, etc. The immediate problem arises with their identification, since they can be written in diverse and inconsistent ways. Take for example the following different spellings which are acceptable in both English and Setswana.

<i>houseboat</i>	<i>house-boat</i>	<i>house boat</i>
<i>tradeoff</i>	<i>trade-off</i>	<i>trade off</i>
<i>khuduthamaga</i>	<i>khudu-thamaga</i>	<i>khudu thamaga</i>
<i>pelotshetlha</i>	<i>pelo-tshetlha</i>	<i>pelo tshetlha</i>
<i>rampatshetlha</i>	<i>rampa-tshetlha</i>	<i>rampa tshetlha</i>
<i>motshwaradiphala</i>	<i>motshwara-diphala</i>	<i>motshwara diphala</i>
<i>kgakalakgakala</i>	<i>kgakala-kgakala</i>	<i>kgakala kgakala</i>

In computational linguistics the examples *houseboat* and *kgakalakgakala* will each constitute a single token, while *house boat* and *kgakala kgakala* will form two different tokens each. Words joined by a hyphen can either be recognised as single words or

as two separate words depending on the tokenising programme. The difference is not trivial in statistical linguistics, since the number of tokens will vary significantly depending on what is counted.

In this chapter we discuss results of three strategies of how the Setswana dictionary headwords could be enriched through the harvesting of concordance data (Sinclair 1991), word association measures and through the generation of concgrams (Scott 2010). Through the three strategies, three broad types of information are revealed. First, there are high frequency words that are found in the vicinity of a word under investigation which are nevertheless not immediately critical to the meaning of a headword. Second, multi-word units (e.g. *pula ya matlakadibe*: a vicious rainy storm, *pula ya sephai*: the first rain of the season; *pula ya kgogolamoko*: the first rain after harvest etc) are identified. Third, a word's valency is revealed. For instance the noun *pula* 'rain' can be followed or preceded by certain Setswana terms such as verbs and adjectives that characterise the type, intensity, end or beginning of the rain. For instance words that express the sense of heavy rain are: '*tsorotla*', '*porotla*', '*bokete*', '*kgolo*', '*tshologa*', '*gosomana*' '*maswe*' and '*tsora*'. '*sarasara*', '*komakoma*' and '*rotha*' all express 'light showers'. '*thiba*' expresses impending rain while '*simolola*', '*itelekela*', and '*kgomoga*' all indicate the start of rain with '*kgomoga*' implying the beginning of a heavy rain or an unexpected rain. '*kgaotsa*', '*didimala*', and '*ema*' relate to the sense of 'stop raining'.

Information relating to Category 1 above may be treated in large Setswana dictionaries. Category 2 "information" is lexicalised and should be included either as independent headwords or as dictionary subentries. Category 3 "collocations" is what could be added as part of a dictionary's usage notes to illustrate the natural collocates of a headword. This will aid users, particularly users of an active dictionary (Svensén 1993) to produce 'natural-sounding' pieces of large units of language. The chapter discusses how these could be harvested to enrich a Setswana dictionary database. The study by Mphahlele (2003) on multiword units has demonstrated that multi-word units are as important as single-word units.

## 2. METHODOLOGY AND EXPERIMENTS

For our experiments, we follow Brunner and Steyner's (2008) approach and use corpus data. By a corpus we mean "a collection of texts, of written or spoken words, which is stored and processed on computer for the purpose of linguistic research" (Renouf 1987:1). The Setswana corpus used for the experiments is just over 15 million tokens. The software employed is *Oxford Wordsmith Tools Version 5* (Scott 2010). It is applied to study a specific word in context in detail in terms of co-texts to its left and to its right.

## 2.1. The Data

The total Setswana corpus compiled is about 15½ million tokens. Ninety four percent of the corpus is the written component while the spoken (transcribed text) component is 6%. Table 1 gives the size of the corpus, on the basis of tokens, type/token ratio (TTR) and standardised type/token ratio (STTR) measures of the whole corpus.

The type/token ratio (TTR) is calculated by dividing types by tokens and multiplying by 100. By types we refer to the *different types* of words that occur in a corpus while by tokens we refer to the count of every word regardless of its repetition. Thus if the word *gore* occurs in a corpus 3145 times, it is said to constitute a single type but 3145 tokens.

The TTR however varies widely in accordance with the length of a text; with shorter texts, the statistic is much more likely to give higher TTR, while longer texts result with a smaller TTR (Malvern and Richards 2002). Because of this phenomenon McKee *et al.* show that TTR measures are flawed,

...because the values obtained are related to the number of words in the sample... samples containing larger numbers of tokens give lower values for TTR and vice versa. ...as longer and longer samples of language are produced, more and more of the active vocabulary is likely to be included and the available pool of *new* word types that can be introduced steadily diminishes. ...it is also the case that however small the sample is, as more and more tokens are taken, the likelihood is that (because of repetition of previously included types) the cumulative number of types will increase at a slower rate than the number of tokens and the TTR values will inevitably fall (McKee *et al.*, 2000: 323).

The solution to this challenge is to compare equal sized text types. The results of comparing Setswana texts would have been much more significant if the text types were of the same size such as in the LOB and Brown Corpus subcorpora (Hofland and Johansson 1982). A more reliable measurement is that of the standardised type/token ratio (STTR).

The ratio for STTR is calculated at every specified number of tokens and an average of the different ratios computed. STTR is computed every *n* words as Wordlist goes through each text file. For the experiments, *n* = 1,000. In other words the ratio is calculated for the first 1,000 running tokens, and then calculated afresh for the next 1,000, and so on to the end of the text or corpus. A running average is computed, which means that we get an average type/token ratio based on consecutive 1,000-word chunks of text. Texts with less than 1,000 words get a standardised type/token ratio of 0. STTR measures are attractive since they can compare

type/token ratios across texts of differing lengths since what they do is segment a corpus into comparable chunks and calculate the type/token ratio for each.

**Table 1: Overall corpus statistics**

File size (bytes)	90,995,080
Tokens	15,053,304
Types	253,525
type/token ratio (TTR)	1.68
standardised TTR (STTR)	34.49

We use the Concord tool and WSConcgram utility of *WordSmith* to run the experiments discussed below. WordSmith Tools is an integrated suite of three main programmes: *Word list*, *Concord* and *Keywords*. The *Word list* tool can be used to produce word lists or word-cluster lists from a text and render the results alphabetically or by frequency order. It can also calculate word spread across a variety of texts, that is, render results on the basis of their spread in different texts. The concordancer, *Concord*, can give any word or phrase in context – so that one can study its co-text; see what other words occur in its vicinity. *KeyWords* calculates words which are key in a text; words which are used much more frequently or much less frequently in a given corpus than expected when measured against a general corpus of the language. For this paper, only frequency and keyword analysis are used.

## 2.2. The Approach

In this study three strategies are employed to identify multiword units. First is the generation of concordance lines, second is the word association measures and finally the concgrams.

## 2.3. Concordance Lines

One way of identifying multi-word units in a corpus is by studying a specific word in context in detail in terms of its co-text to the left and to its right. This is achieved by generating a key word in context (KWIC) often referred to as concordance lines. "A concordance is an index of the surface word forms in a text. It is a collection of the occurrences of a word form, each in its own textual environment." (Dash and Chaudhuri 2000:190). A concordance reveals the company kept by a word, its collocates, and thereby reveals meanings and usages which are hard to dig up through mental



recall (Otlogetswe 2007:56). We illustrate this later in the development of the chapter through the example of the tokens *pula* and *tshwara*.

The *pula* token is used to represent a variety of homographic words with different meanings and usages in Setswana. These are illustrated in the following three examples:

- a. It is used to mean “rain” as in the sentence ***Maru a pula a lebega a tlaa anama le lefatshe lotlhe*** (It appears that rain clouds will spread throughout the country.)
- b. *Pula* is also used as a cry of well wishing as in the sentence *Pula! Batho betsho!* (Pula! Fellow people!)
- c. *Pula* is also used as Botswana’s currency as in the sentence *Ba batla o duela ka Pula fela*. (They want you to pay in Pula only.)

Below is an example of concordance lines for the word *pula*.

a. Loapi lo rile thi ke maru a pula ya matla-ka-dibe, e bile tla thoa fela. Maloba jale l~a pula ya ntlha e kgolo go kile rosela mo go ene jaaka metsi a pula ya sephai a ela ka mokgo aya tsebe gee, molekane Maru a pula ya matlakadibe a mantsho enya go tswa bophirima, maru a pula ya Kgalagatsana. Maikael osa tsa tshi-ngwana ya Gago. A pula ya Gago e ba nele; a Let di tshwantshanngwang le maru a pula ya tsheola. Pula e e tsh ; E phepa jaaka metsi, metsi a pula ya sephai. Moso le moso mo matlhong jaaka marothodi a pula ya sefako e na ka diphef la Re tie re tshela ka metsi a pula ya gago Ditshaba di dike ne a ne a gogopelwa ke metsi a pula ya kgogolammoko mo lewat polotiki Wa ikala ja~ka maru a pula ya medupe Mokgosi wa utl kwa holong ya sekolo sa Maru-A-Pula ya Maitisong morago ga g gobo ~E tswana le marothodi a pula ya medupe, Motswedi wa b e (b) SerampheetAhane 158. (a) Pula ya na matsorotsoro, nna a ba ratela fela go ba roma, a pula ya na kgotsa letsatsi le o o lebetse eng? Monna, maru a pula ya matlakadibe a a kokom e e fa godimo e? b Marothodi a pula ya medupe a ka oketsa bo gobo ~E tswana le marothodi a pula ya medupe, Motswedi wa b polotiki Wa ikala ja~ka maru a pula ya medupe Mokgosi wa utl a ba ratela fela go ba roma, a pula ya na kgotsa letsatsi le enya go tswa bophirima, maru a pula ya Kgalagatsana. Maikael tla thoa fela. Maloba jale l{a pula ya ntlha e kgolo go kile sheka jaaka seretse sa metse a pula ya sephai mogobing wa Ta

To identify a word’s collocates one can study its patterns, that is, words that typically occur on its left (L) or on its right (R) in high frequency. It is by inspecting collocates that we can uncover different MWEs such as proverbs, compounds, idioms, sayings, phrasal verbs, etc. Such structures can then be entered into dictionaries as subentries, multi-word headwords or collocates. Through the use of computer programmes or concordance software, it is relatively easy to obtain a list of all the co-occurrences of a particular word in context and see all the meanings associated with such a word (Biber *et al.* 1998: 27). The concordance lines above reveal the different subtle meanings associated with the word *pula*. For instance in

the above data, words such as *medupe*, *matlakadibe*, *kgogolamoko*, *tsheola*, *sefako* are found on the R2 position of the search word.

There are however 3884 concordance lines for the *pula* search word which would be too much for individual inspection. The collocates can however be summarised as shown below in what Wordsmith calls *patterns*. This shows the collocates (words adjacent to the search word), organised in terms of frequency within each column. That is, the top word in each column is the word most frequently found in that position. The second word is the second most frequent, etc.

L4	L3	L2	L1	centre	R1	R2	R3	R4
a	a	go	a	pula	e	e	e	a
e	go	a	fa		ya	a	a	e
go	e	e	ke		a	ne	go	go
ka	ba	le	ka		pula	na	ya	le
le	le	maru	nesa		o	nele	le	na
ba	ka	ya	ga		le	ka	ka	ba
o	ke	ba	wa		go	le	na	ka
ya	ya	ka	pula		mo	tla	ba	o
ke	di	ga	la		ke	go	ne	sa
re	mo	re	ya		ka	medupe	ga	ya
mo	o	metsi	gore		ga	sa	o	mo
fa	re	morago	le		ba	ke	mo	ke
ga	ne	se	re		mme	o	ke	re
ne	fa	mosele	jaaka		di	re	re	ne
di	ga	o	nesetsa		fa	ba	pula	ga
se	pula	ledi	na		tse	pula	fa	di
pula	jaaka	marothodi	tsa		mosi	matlakadibe	mme	se
gore	sa	pula	se		kwa	ya	se	pula
wa	gore	ke	sa		re	ga	kgotsa	fa
mme	maru	mo	mme		eo	kgolo	tla	bo

The results demonstrate the following:

- The most frequent L1 words are: *a*, *fa*, *ke*, *ka*, *nesa*, *ga*, *wa*, *pula*, *la*, *ya*, *gore*, *re*, *jaaka*, *nesetsa*, *na*, *tsa*, *se*, *sa*, *mme*, *bona*, *di*, *teng*, *kopa*, *rapelela*, *tshabela*, *ralala*, and *tshologa*.
- The most frequent L2 words are: *go*, *a*, *e*, *le*, *maru*, *ya*, *ba*, *ka*, *ga*, *re*, *metsi*, *morago*, *se*, *mosele*, *o*, *ledi*, *marothodi*, *pula*, *ke*, *mo*, *na*, *di*, *nelwa*, *nelwe*, *wa*, *sa*, *tsa*, *kgole*, *leru*, *newa*, and *gosomana*.
- The most frequent R1 words are: *e*, *ya*, *a*, *pula*, *o*, *le*, *go*, *mo*, *ke*, *ka*, *ga*, *ba*, *mme*, *di*, *fa*, *tse*, *mosi*, *kwa*, *re*, *eo*, *tsa*, *nkgodisa*, and *bagaetsho*.
- The most frequent R2 words are: *e*, *a*, *ne*, *na*, *nele*, *ka*, *le*, *tla*, *go*, *medupe*, *sa*, *ke*, *o*, *re*, *ba*, *pula*, *matlakadibe*, *ya*, *ga*, *kgolo*, *sephai*, *sefako*, *simolola*, *tsheola*, *tsamaya*, *namagadi*, *selemo*, *maebana*, *dikgadima*, *ntsi*, and *ditladi*.

It is from the inspection of terms at L1, L2, R2, and R1 positions that we identify words that collocate with *pula*. These terms can then be culled and used to enhance a dictionary database. We demonstrate how this could be implemented later in the development of this paper.

The next analysis is of the verb *tshwara* which contextually could mean touch, handle or hold either physical or mentally.

e go di mmaya pele, a tsena mo ntiong. 0 tshwara sele, a tiogela fa, a sutisa se  
sa tshwanela go tewa, ke mongwe gore 0' tshwara semangmang, me semangmang ene  
ngwe jaanong 0 kare 0 tlaa tlola metsi 0 tshwara seretse. Rraetsho, nna ke sa nt  
kanyo, kwa ntle ga motsetsi mmangwana 0 tshwara thipa ka fa bogaleng. Ba feta K  
thale o, mme Badir-leng e bong mmaugwe 0 tshwara thipa ka fa bogaleng. Rrangwan'  
hata a tsee sesagagwe! Mmangwana ke yo 0 tshwara thipa ka bogale! Dietapele di a  
menate va torono bomolekane v Mosimane 1 Tshwara molekane. Re go bonetse tshwabi  
abanong ka phata ya motse wa Herese, 14a tshwara lekau la kwa Sukothe, a le bots  
21:2. s) Baatlh. 1:16. 362 I SAMUELE 15 tshwara Agage kgosi ya Baamaleke a sa n  
timatima 0 tingwa ke phefo e' bogale 180 Tshwara 'tomo morwa Masire-a-Ketumile M  
eleng kafa tlase ga maemo a a boima. 2. Tshwara batho ka tolamo, ka maitseo le  
lete weel!...A ke fatlhwa ke letsatsi? 5 Tshwara ka thata ngwana wa ga Seboko Se  
30:27. r, Diane s 3. u) Ekes. 20:14. 50 tshwara ka seaparo sa gagwe a re: "Roba  
wa go Eskom boikaelelo jwa yone fela. 7. Tshwara didirisiwa le dithoto tsa Eskom  
ng tse di tshesane gonne tsona di tla 71 tshwara molelo ka pele. Fa masasana a s  
o (c) 8atho ba kile ba suga sekgoropa 8a tshwara tlalo la kgomo, morwa-Kgama La

*Tshwara* is more frequent than *pula* in the Setswana corpus. Its analysis generated 5,340 concordance lines. However, such massive data may be summarised in the following manner:

L4	L3	L2	L1	centre	R1	R2	R3	R4
a	a	a	go	tshwara	ka	ka	a	a
ba	ba	go	a		a	a	e	go
go	o	ba	mo		ba	le	go	e
o	go	ka	ba		bothata	ba	ba	mo
ka	le	ke	e		fa	e	o	ba
e	ke	ne	ka		thipa	ya	le	ka
fa	re	ya	o		ditlhapi	mo	fa	le
le	ka	tla	di		e	go	ka	ya
re	fa	o	tla		o	fa	di	o
mo	e	re	re		mo	o	mo	fa
ya	ya	le	le		ke	wa	ke	ke
ke	se	fa	sa		motho	tse	ya	ne
ga	gore	mme	se		kwa	tse	se	re
gore	ne	mmangwana	ke		sentle	la	re	ga
ne	sa	di	ya		fela	ke	gagwe	se
se	tla	e	wa		le	mme	ga	bogaleng
di	mo	tse	ra		mme	seatla	kwa	tse
sa	di	kgona	o		sepe	diatla	ne	mme
kwa	ga	bo	bo		batho	sa	tse	di
wa	bo	wa	lo		tiro	ga	bo	kwa
mme	mme	sa	la		go	se	seatla	wa

itse	gagwe	leka	tsa	legodu	letsogo	wa	sa
fela	leka	la	gago	diphuthego	re	lesa	gore

The results of the concordance study of *tshwara* show that *tshwara* collocates with the following words in the stated positions.

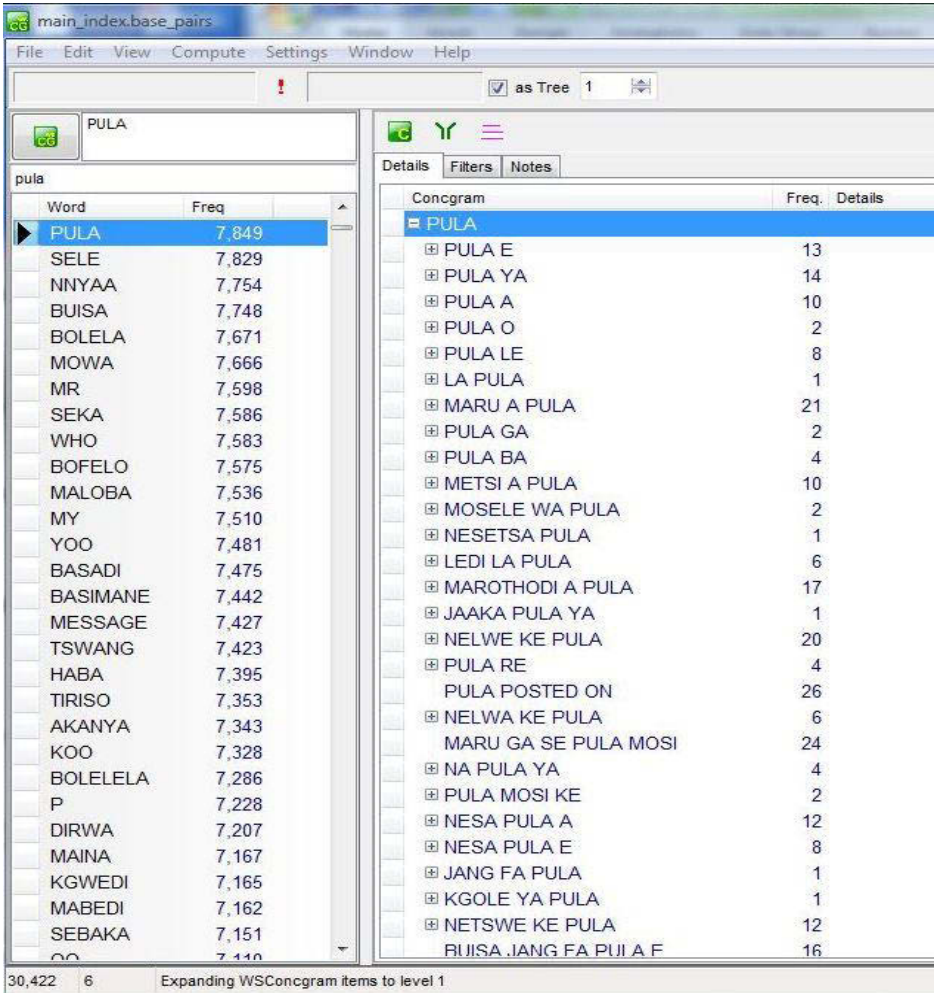
- a. The most frequent R1 words are: *ka, a, ba, bothata, fa, thipa, ditlhapi, e, o, mo, ke, motho, kwa, sentle, fela, le, mme, sepe, batho, tiro, go, legodu, diphuthego, letsogo, thata, pelo, terena, masome, mogala, bese, kgomo, kgole, boroko, phuthego, legaba, ntlha, phage, dithuto, dipuisano, pitso, mafoko, botlhaswa, kgaisano, mala, and sejana.*
- b. The most frequent R2 words are: *ka, a, le, ba, e, ya, mo, go, fa, o, wa, tsa, tse, la, ke, mme, seatla, diatla, sa, ga, se, letsogo, re, thata, sentle, lesa, tlhogo, rona, botlhe, natla, tseleng, mongwe, lebogo, thamo, tlhaa, legetla, mmetsa, pelo, and ditsebe.*
- c. The most frequent L1 words are: *go, a, mo, ba, e, ka, o, di, tla, re, le, sa, se, ke, ya, wa, ra, bo, lo, la, tsa, gago, tshwara, nka, buisa, tlaa, wena, and gore.*
- d. The most frequent L2 words are: *a, go, ba, ka, ke, ne, ya, tla, o, re, le, fa, mme, mmangwana, di, e, tsa, kgona, bo, wa, sa, leka, la, tshwanetse, batla, nka, sena, rata, sena, fitlha, tlola, bona, kgonang, and batho.*

The words which collocate with *pula* and *tshwara* provide a rich variety of material which may be used as additional information about word use and correlation. How the collocates may be used in a dictionary, is discussed later in the development of this paper. The next section introduces another strategy of identifying multiword units through the study of concgrams.

## 2.4. Concgrams

A ‘concgram’ has been defined as all of the permutations of constituency variation and positional variation generated by the association of two or more words (Cheng *et al.* 2008). This means that the associated words comprising a particular concgram may be the source of a number of ‘collocational patterns’ (Sinclair 2004:xxvii). Scott (2010) has argued that what is termed ‘concgrams’ has a long history dating back to the 1980s when the COBUILD team at the University of Birmingham, led by John Sinclair, attempted to devise means to automatically search for non-contiguous sequences of associated words (Cheng *et al.* 2006:414 in Scott 2009:240). Cheng *et al.* (2006:415) suggest that concgrams are a starting point for identifying and quantifying the extent of phraseology in a text or corpus, and hence the phraseological profile of the language contained within it (Cheng *et al.* 2008:236). The concgram measures for the word *pula* yielded the following results:

Figure 1: The concngram measures for the word *pula*



The results of the analysis demonstrate that statistically, the following are found after 'pula e'

*ne, na, nele, tla, kgolo, namagadi, ntsi, tshweu, tswa, boutsana, simolotse, tona, kgaotsa, bokete, phaila, rotha, tsorotla, utlwala, duma, goroga.*

Studying *pula* with its concords is important since Setswana nouns, as in other Bantu languages, are usually followed by various concords in a sentence. *Pula* followed by the noun class 4 possessive concord *ya* takes the following arguments:

*medupe, matlakadibe, na, sefako, ntlha, tsheola, dikgadima, selemo, tla, sephai, kgogolamoko, tshologa, simolola, leebana, mariga, matlotlo, morago, morwalela, ngwaga, ditladi, maebana*

The following MWEs have also been extracted statistically through concgram measures:

*maru a pula, nesa pula, metsi a pula, morago ga pula, mosele wa pula, nesetsa pula, marothodi a pula, ledi la pula, nelwa ke pula, nelwe ke pula, rapelela pula, ralala pula, gosomana ga pula, bokete jwa pula, reka ka pula, kolobeditswe ke pula, tshwarwa ke pula, leru la pula, lerothodi la pula, go na pula, rapelela pula, setlha sa pula, ya tsholola pula, modumo wa pula, mookodi wa pula, moroka wa pula, kgole ya pula*

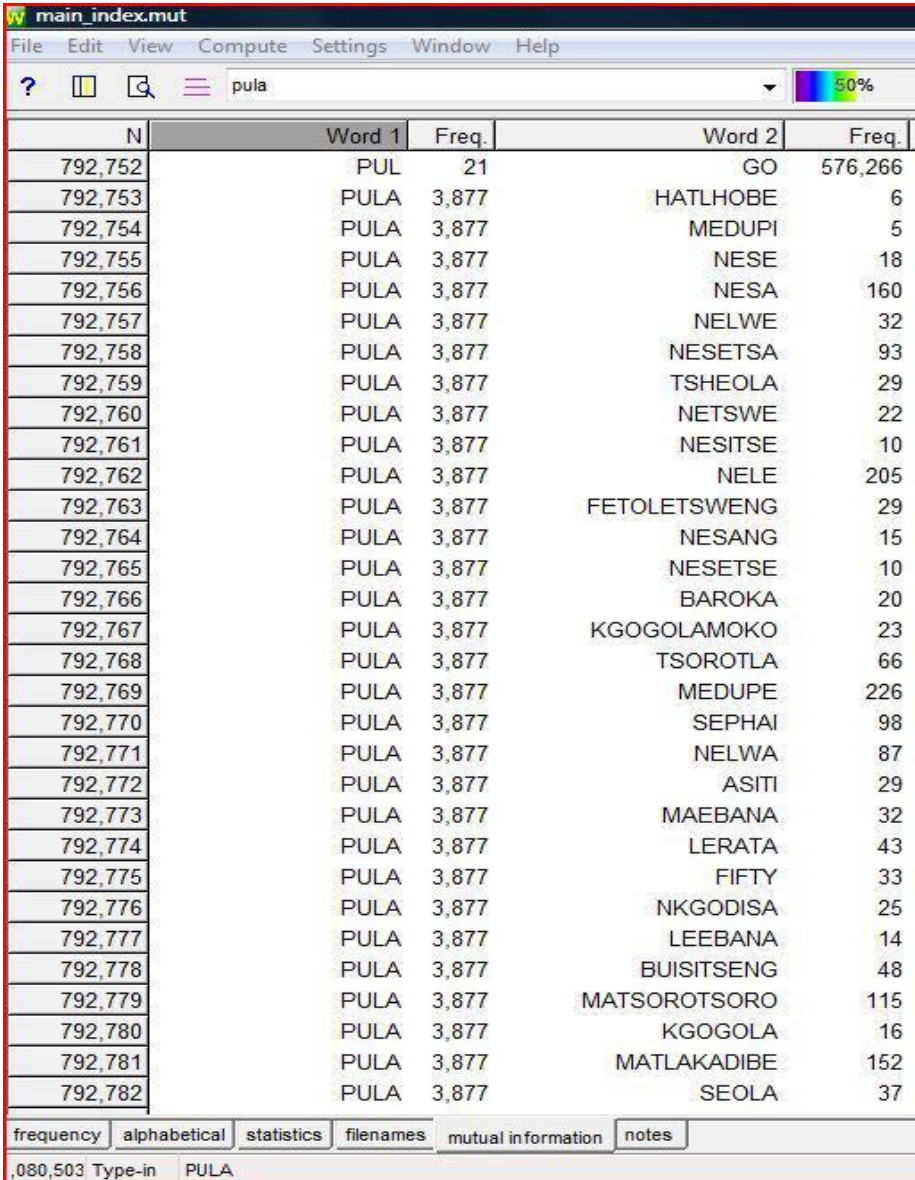
These results are critical since they demonstrate that multi-word units and idiom fragments can be extracted automatically through the Concgram programme fairly cheaply. These are potential headwords, subentries and collocates.

Finally, we consider mutual information measures in identification of multiword units.

### 3. MUTUAL INFORMATION MEASURES

The Mutual Information (MI) measure in its application to lexicography was first popularised by Church and Hanks who argued that results of MI measures were relevant to lexicography in “enhancing the productivity of lexicographers in identifying normal and conventional usage” (Church and Hanks 1989:76). Additional to their argument, is that MI measures are able to isolate the word’s collocates and possible multiword expressions which can be used to enhance a dictionary headword list. Mutual Information (MI) score relates one word to another. For example, if *problem* is often found with *solve*, they may have a high mutual information score. Usually, *the* will be found much more often near *problem* than *solve*, so the procedure for calculating Mutual Information takes into account not just the most frequent words found near the word in question, but also whether each word is often found elsewhere, away from the word in question. Since *the* is found very often indeed far away from *problem*, it will not tend to be related, that is, it will get a low MI score (Scott 2010:199). Mutual information can help us decide what to look for in the concordance since it provides a quick summary of what company words keep (Firth 1957).

Our experiment measures the likelihood to which words are likely to occur next to the noun *pula* and the verb *tshwara*. The results of the experiment are given in Figures 2 and 3.

Figure 2: MI measures for the word *pula*


The screenshot shows a software window titled 'main\_index.mut' with a menu bar (File, Edit, View, Compute, Settings, Window, Help) and a search bar containing 'pula'. Below the search bar is a table with columns: N, Word 1, Freq., Word 2, and Freq. The table lists 30 rows of data, each representing a word associated with 'pula' and its frequency. At the bottom of the window, there are tabs for 'frequency', 'alphabetical', 'statistics', 'filenames', 'mutual information', and 'notes'. The 'mutual information' tab is currently selected, showing the data table. Below the table, it displays ',080,503 Type-in' and 'PULA'.

N	Word 1	Freq.	Word 2	Freq.
792,752	PUL	21	GO	576,266
792,753	PULA	3,877	HATLHOBE	6
792,754	PULA	3,877	MEDUPI	5
792,755	PULA	3,877	NESE	18
792,756	PULA	3,877	NESA	160
792,757	PULA	3,877	NELWE	32
792,758	PULA	3,877	NESETSA	93
792,759	PULA	3,877	TSHEOLA	29
792,760	PULA	3,877	NETSWE	22
792,761	PULA	3,877	NESITSE	10
792,762	PULA	3,877	NELE	205
792,763	PULA	3,877	FETOLETSWENG	29
792,764	PULA	3,877	NESANG	15
792,765	PULA	3,877	NESETSE	10
792,766	PULA	3,877	BAROKA	20
792,767	PULA	3,877	KGOGOLAMOKO	23
792,768	PULA	3,877	TSOROTLA	66
792,769	PULA	3,877	MEDUPE	226
792,770	PULA	3,877	SEPHAI	98
792,771	PULA	3,877	NELWA	87
792,772	PULA	3,877	ASITI	29
792,773	PULA	3,877	MAEBANA	32
792,774	PULA	3,877	LERATA	43
792,775	PULA	3,877	FIFTY	33
792,776	PULA	3,877	NKGODISA	25
792,777	PULA	3,877	LEEBANA	14
792,778	PULA	3,877	BUISITSENG	48
792,779	PULA	3,877	MATSOROTSORO	115
792,780	PULA	3,877	KGOGOLA	16
792,781	PULA	3,877	MATLAKADIBE	152
792,782	PULA	3,877	SEOLA	37

frequency alphabetical statistics filenames mutual information notes  
,080,503 Type-in PULA

The findings of this study reveal that the following words have a high association with the word *pula*: *go, medupi, nese, nese, nelwe, nesetsa, tsheola, netswe, nesitse, nele, fetoletsweng, nesang, nesetse, baroka, kgogolamoko, tsorotla, medupe, sephai, nelwa, asiti, maebana, lerata, fifty, nkgodisa, leebana, buisitseng, matsorotsoro, kgogola, matlakadibe, seola*.

The results for *tshwara* follow below.

**Figure 3: MI measures for the word *tshwara***

N	Word 1	Freq.	Word 2	Freq.
1,023,536	TSHWARA	5,308	SERETSE	834
1,023,537	TSHWARA	5,308	MAREKO	1,337
1,023,538	TSHWARA	5,308	KGONANG	1,231
1,023,539	TSHWARA	5,308	BOMONNAWE	338
1,023,540	TSHWARA	5,308	TLOLA	1,979
1,023,541	TSHWARA	5,308	MAPODISA	454
1,023,542	TSHWARA	5,308	KGAKGE	287
1,023,543	TSHWARA	5,308	PALELWA	2,020
1,023,544	TSHWARA	5,308	MAGODU	638
1,023,545	TSHWARA	5,308	KGETSE	1,185
1,023,546	TSHWARA	5,308	LEOTO	712
1,023,547	TSHWARA	5,308	RREMOGOLO	476
1,023,548	TSHWARA	5,308	TLOU	1,196
1,023,549	TSHWARA	5,308	MEDI	359
1,023,550	TSHWARA	5,308	SEFOFANE	423
1,023,551	TSHWARA	5,308	LELE TSA	373
1,023,552	TSHWARA	5,308	FALE	1,990
1,023,553	TSHWARA	5,308	MAKGABE	376
1,023,554	TSHWARA	5,308	DINTSWA	378
1,023,555	TSHWARA	5,308	DILE	380
1,023,556	TSHWARA	5,308	LOGAGA	381
1,023,557	TSHWARA	5,308	ATAMELA	1,975
1,023,558	TSHWARA	5,308	DIBESE	322
1,023,559	TSHWARA	5,308	LEANO	1,302
1,023,560	TSHWARA	5,308	MENWANA	1,174
1,023,561	TSHWARA	5,308	SENNA	397
1,023,562	TSHWARA	5,308	DUELWA	399
1,023,563	TSHWARA	5,308	TAU	3,328
1,023,564	TSHWARA	5,308	IKAELELA	1,294
1,023,565	TSHWARA	5,308	MAGODIMO	412
1,023,566	TSHWARA	5,308	LOGA	689

frequency alphabetical statistics filenames mutual information notes

1,080,503 Type-in TSHWARA

Words with the greatest association to *tshwara* include: *seretse*, *mareko*, *kgonang*, *bomonnowe*, *tlola*, *mapodisa*, *kgakge*, *palelwa*, *magodu*, *kgetse*, *leoto*, *rremogolo*, *tlou*, *medi*, *sefofane*, *leletsa*, *fale*, *makgabe*, *dintswa*, *dile*, *logaga*, *atamela*, *dibese*, *leano*, *menwana*, *senna*, *duelwa*, *tau*, *ikaelela* *magodimo* *loga*, etc.



So far concgrams, mutual information measures and concordances have been used to identify word associations. Collocates that have been extracted may be classified into three broad groups for dictionary consideration. These are:

### 3.1. Lexicographically Uninteresting/Minimally Interesting Token Associations

These include various concords that associate with the search word. As with many members of the open word classes, these concords associate highly with *tshwara* and *pula*, but also with many other words in the language, and thus not making their association statistically and lexicographically interesting. Such associations are captured in the following evidence:

#### **tshwara**

- |               |               |                |
|---------------|---------------|----------------|
| a. tshwara ka | e. tshwara mo | i. tshwara tse |
| b. tshwara a  | f. tshwara o  | j. tshwara ga  |
| c. tshwara ba | g. tshwara re | k. tshwara go  |
| d. tshwara e  | h. tshwara fa |                |

#### **pula**

- |            |             |            |
|------------|-------------|------------|
| a. pula e  | f. pula kwa | k. pula go |
| b. pula ya | g. pula re  | l. pula ke |
| c. pula a  | h. pula eo  | m. pula re |
| d. pula o  | i. pula tsa |            |
| e. pula di | j. pula a   |            |

### 3.2. Headwords/Subentry Candidates

The second class is of those clusters which are lexicographically interesting. These include idioms, proverbs, and multiword names which may be added to the dictionary either as independent headwords or as dictionary subentries. These are particularly interesting since they function as independent lexical items because of their semantically non-compositional nature. When idiomatic collocates are treated as subentries in dictionaries or as independent headwords, it is important that the type of dictionary should be kept in mind. Normally general dictionaries, which have a more inclusive nature, can accommodate more subentries than standard or smaller school dictionaries, which, because of their smaller size, have to exclude many sub-entries. In the case of very economical, restrictive and selective dictionaries, all subentries will have to be omitted.

Therefore when, in the following discussion, we indicate how the subentries in some Setswana dictionaries may be increased, it does not necessarily mean that all

these subentries should be included in every Setswana dictionary. It merely shows what is possible. When a choice has to be made according to the nature of the dictionary, which subentries have to be included in a specific type of dictionary, corpus evidence will be helpful to indicate which idiomatic collocates are the most commonly and generally used. Below are some of the MWEs which could be entered in a Setswana dictionary as either subentries, independent headwords or as collocates.

### pula

maru ga se pula, mosi ke molelo nesa ke pula	Suspicion usually have a basis
mosele wa pula o etšwa go sa le gale	Make hay while the sun shines
kgole ya pula e bošwa e bofologa	You can discipline someone, but they are bound to deviate from their instruction
pula ya medupe	A quiet rain that falls for a long time
pula ya sephai	The first rain of the season
pula ya maebana	A good rain that falls quietly for a long time
pula ya tsheola	A rain that falls after ploughing
pula ya kgogolamoko	The first rain after harvest
pula e namagadi	Good rain
pula e tshweu	Good rain without lightning and thunder

### tshwara

tshwara ditlhapi	To fish
tshwara pelo	To control one's emotions
tshwara bothata	To face difficulties
tshwara logaba	To eat a little especially when very hungry
tshwara thipa ka fa bogaleng	To defend someone that you like
Tshwara phage ka magana	To face problems
tshwara poo	To take someone's belongings (by thugs)
tshwara mala ka letsogo	To be scared

### 3.3. Word-Associations Which Would Enhance Usage Boxes

The third class is that of words which collocate with the search word. These words are important in helping a student or a writer produce natural, native-speaker like writing. These words could be included in collocation boxes as done in the *Macmillan English Dictionary for Advanced Learners* (Rundell 2007). An impressive result of lexicographic application of collocation study is the *Oxford Collocations Dictionary for students of English* (Lea *et al.* 2002) which includes over 150,000 collocations of 9,000 nouns, verbs and adjectives with over 50,000 examples of collocations in context. What follows are words which collocate with *pula* and *tshwara* which could be used in collocation boxes of a dictionary.

#### **pula**

*na, nele, tla, kgolo, namagadi, ntsi, tshologa, tshweu, tswa, boutsana, simolotse, tona, kgaotsa, bokete, phaila, porotla, rotha, tsorotla, utlwala, duma, goroga, ntlha, tsheola, dikgadima, selemo, tla, tshologa, simolola, mariga, matlotlo, morago, morwalela, ngwaga, ditladi.*

#### **tshwara**

*bothata, sentle, sepe, diphuthego, legodu, letsogo, thata, terena, boroko, phuthego, ntlha, phage, dithuto, dipuisano, pitso, mafoko, botlhaswa, kgaisano.*

## 4. REVIEWING DICTIONARY ENTRIES

Setswana dictionaries have attempted to include subentries based on the idiomaticity of collocates. However, subentries are few in Setswana dictionaries often because of a lack of sufficient corpus evidence. Above, through the study of *pula* and *tshwara*, we have demonstrated what could be added to a dictionary to enhance its functionality. Presented below are examples of how the entries *pula* and *tshwara* have been treated in Kgasa and Tsonope (1998).

These treatments are followed by suggested revisions of the headwords with the lexical wealth identified above.

Treatment of *pula* as entry in Kgasa and Tsonope (1998: 225):

**pula** GT In./9. di- 1. metsi a a tshologang go tswa kwa godimo mo marung. 2. madi a lefatshe la Botswana. \* *pula ga e ke e be e swela godimo = le fa go ka nna leuba la dingwagangwaga, pula e tle e be e ne.*

This entry may be revised in the following manner:

**pula** GT *In./9. di-* 1. metsi a a tshologang go tswa kwa godimo mo marung. 2. madi a lefatshe la Botswana. ■ *mosele wa pula o etšhwa go sa le gale.* ■ *Nelwa ke pula* ■ *Maru ga se pula mosi ke molelo* ■ *Kgole e boswa e bofologa* ■ *Pula ya medupe* ■ *Pula ya sephai* ■ *Pula ya maebana* ■ *Pula e tshweu* ■ *Pula ya tsheola* ■ *Pula ya kgogolamoko* ■ *Pula e namagadi.*

Matlamorago ga “ <b>pula e</b> ”
na, nele, tla, kgolo, namagadi, ntsi, tshologa, tshweu, tswa, boutsana, simolotse, tona, kgaotsa, bokete, phaila, porotla, rotha, tsorotla, utlwala, duma, goroga.
Matlamorago ga “ <b>pula ya</b> ”
Medupe, matlakadibe, sefako, ntlha, tsheola, dikgadima, selemo, tla, sephai, kgogolamoko, tshologa, simolola, leebana, mariga, matlotlo, morago, morwalela, ngwaga, ditladi, maebana.

Treatment of *tshwara* as entry in Kgasa and Tsonope (1998:225):

**tshwara** GT *tpt* -ile 1. baya seatla mo sengweng 2. mogapa (sic = gapa) sengwe gore se nne mo tlhokomelong ya gago 3. tshola sengwe ka mokgwa mongwe \* *go tshwara matlhaku ka dithito = go leka ka thata.*

This entry may be revised in the following manner:

**tshwara** GT *tpt* -ile 1. baya seatla mo sengweng 2. mogapa (sic = gapa) sengwe gore se nne mo tlhokomelong ya gago 3. tshola sengwe ka mokgwa mongwe ■ *go tshwara matlhaku ka dithito* ■ *mmangwana o tshwara thipa ka fa bogaleng* ■ *tshwara tshwara* ■ *tshwara ditlhapi* ■ *tshwara logaba* ■ *tshwara phage ka mangana* ■ *tshwara mala ka letsogo* ■ *Tshwara pelo* ■ *tshwara pelo* ■ *mmangwana o tshwara thipa ka fa bogaleng* ■ *Tshwara poo*

<i>Matlamorago ga "tshwara"</i>
bothata, ditlhapi, sentle, sepe, diphuthego, legodu, letsogo, thata, pelo, terena, kgole, boroko, phuthego, legaba, ntlha, phage, dithuto, dipuisano, pitso, mafoko, botlhaswa, kgaisano, mala.

The proposed revisions demonstrate what is possible if the results of word associations study were to be used to enrich lexical databases for lexicography, in particular, lexicography in African languages. While research in English and other European languages has benefited immensely from statistical analysis and the study of word associations, African languages are only discovering the power of lexical computing. African languages have a rich collection of idioms, proverbs and sayings which can be extracted accurately from a corpus.

## 5. CONCLUSION

In this chapter, we have attempted to illustrate what could be achieved by a study of concordance lines to extract MWEs for the significant improvement of dictionary entries. Considering only single words as candidates for dictionary entry impoverishes a dictionary and betrays a rudimentary understanding of what constitutes a word in language.

If Jackendoff's (1997) estimate that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words is accurate, then MWEs in African languages deserve intensive study, which they have hitherto not received. To generate concordance lines is inexpensive, and free concordance programmes such as the Simple Concordance program<sup>1</sup>, TextSTAT<sup>2</sup>, AntConc<sup>3</sup> and Adelaide Text Analysis Tool<sup>4</sup> are available online to aid researchers explore the complexity of texts.

Dictionaries of African languages would therefore benefit greatly from populating subentries with MWEs harvested from concordance lines. MI measures as well as concgrams have also demonstrated their ability to aid in the extraction and identification of word associations which could then be used by lexicographers to enrich the dictionary either as subentries, headwords or collocations. Since the corpus used in this study is not tagged, the study shows that although it is tedious, it is possible to extract collocations from untagged corpora (Taljad and de Schryver 2002).

A tagged corpus would be better and once a Setswana corpus is tagged, it will be possible to significantly speed up "accurate" extraction of collocations through

such software programmes as Xtract (Smadja 1993) and SketchEngine (Kilgariff and Tugwell 2001).

## ENDNOTES

1. <http://www.textworld.com/scp>
2. <http://neon.niederlandistik.fu-berlin.de/en/textstat/>
3. <http://www.antlab.sci.waseda.ac.jp/software.html>
4. <http://www.adelaide.edu.au/red/adtat/>

## REFERENCES

- Aitchison, J. 1992. *Teach Yourself Linguistics*. London: Hodder & Stoughton.
- Bannard, C. 2007. A Measure of Syntactic Flexibility for Automatically Identifying Multi-word Expressions in Corpora. *Proceedings of the ACL Workshop on a Broader Perspective on Multi-word Expressions*, Prague, Czech Republic, June 2007: 1-8.
- Béjoint, H. 2010. *The lexicography of English*. OUP. Oxford
- Brunner, A. & K. Steyner. 2008. Corpus-Driven Study of Multi-Word Expressions Based on Collocations from a Very Large Corpus. Paper presented at the *Fourth Inter-Varietal Applied Corpus Studies (IVACS) Conference*, University of Limerick, Ireland, 13–14 June 2008.
- Cheng, W., C. Greaves J. Sinclair & M. Warren. 2008. Uncovering the extent of the phraseological tendency: towards a systematic analysis of concgrams. *Applied Linguistics* 3(2):236-252.
- Cheng, W., C. Greaves & M. Warren, 2006, From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11(4):411-433.
- Church, K. W. and Hanks, P. 1989. Word association norms, mutual information and lexicography. *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*. Association for Computational Linguistics.76-83.
- Cowie, A.P. 1998. *Phraseology: Theory, analysis and applications*. Oxford University Press. Oxford.
- Dash, N.S. & B.B. Chaudhuri. 2000. The Process of Designing a Multidisciplinary Monolingual Sample Corpus. *International Journal of Corpus Linguistics* 5(2):179-197.

- Fazly, A. & S. Stevenson. 2007. Distinguishing Subtypes of Multiword Expressions Using Linguistically-motivated Statistical Measures. *Proceedings of the ACL Workshop on a Broader Perspective on Multi-word Expressions*, Prague, Czech Republic.9-16.
- Finch, G. 2000. *Linguistic Terms and Concepts*. Basingstoke: Macmillan Press.
- Firth, J. 1957. A Synopsis of Linguistic Theory 1930-1955. *Studies in Linguistic Analysis*. Oxford: Philological Society. Reprinted in *Selected Papers of J.R. Firth*, edited by F. Palmer (1968). Harlow: Longman.
- Grefenstette, G. 2008. The future of linguistics and lexicography: will there be lexicographers in the year 3000? *Practical Lexicography: a reader*, edited by T. Fontenelle. Oxford: Oxford University Press.
- Hofland, K. & S. Johansson. 1982. *Word frequencies in British and American English*. The Norwegian Computing Centre for the Humanities, Bergen, Norway.
- Jackendoff, R. 1997. *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.
- Kgasa, M.L.A. & J. Tsonope. 1998. *Tharodi ya Setswana*. Gaborone: Longman.
- Kilgariff, A. & D. Tugwell. 2001. Word Sketch: Extraction and Display of Significant Collocations for Lexicography. *Proceedings ACL workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation*. Toulouse, France.32-38.
- Lea, D., J. Crowther & S. Dignen. 2002. *Oxford Collocations dictionary for students of English*. OUP. Oxford.
- Leech, G., M. Deuchar & R. Hoogenraad. 1982. *English Grammar for Today: A New Introduction*. Basingstoke: Macmillan Press.
- Malvern, R. & B. Richards. 2002. Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing* 19(1):85-104.
- McKee, G., D. Malvern & B. Richards. 2000. Measuring Vocabulary Diversity Using Dedicated Software. *Literary and Linguistic Computing* 15(3):323-337.
- Moon, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford: Oxford University Press.
- Mphahlele, M.C. 2003. The lexicographic treatment of sublexical and multilexical items in a Northern Sotho monolingual dictionary: a challenge for lexicographers. *Lexikos* 13:154-167

- Oflazer, K. and Ö. Çetinoğlu. 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. *Second ACL Workshop on Multi-word Expressions: Integrating Processing*, Barcelona, Spain.64-71.
- Otlogetswe, T.J. 2007. *Corpus Design for Setswana Lexicography*. Unpublished Ph.D. Thesis. Pretoria: University of Pretoria.
- Otlogetswe, T.J. 2009. Populating subentries in dictionaries with multi-word units from concordance lines. *Lexikos* 19:446-457.
- Pearsall, J. 1998. *The New Oxford Dictionary of English*. Oxford: Oxford University Press.
- Renouf, A. 1987. Corpus Development. *Looking Up. An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*, edited by J.M. Sinclair. London/Glasgow: COBUILD.
- Sag, I.A., T. Baldwin, F. Bond, A. Copestake & D. Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexico.1-15.
- Schone, P. & D. Jurafsky. 2001. Is Knowledge-free Induction of Multi-word Unit Dictionary Headwords a Solved Problem? *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, Pittsburgh, PA. 100-108.
- Scott, M. 2010. *Oxford WordSmith Tools Version 5*. Oxford University Press. Oxford.
- Sharoff, S. 2004. What is at Stake: A Case Study of Russian Expressions Starting with a Preposition. *Second ACL Workshop on Multi-word Expressions: Integrating Processing*. Barcelona, Spain.17-23.
- Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Smadja, F. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143-177.
- Svensén, B. 1993. *Practical lexicography: Principles and methods of dictionary-making*. Oxford and New York: Oxford University Press.
- Svensén, B. 2009. *A handbook of lexicography: The theory and practice of dictionary making*. Cambridge: Cambridge University Press.
- Summers, D. 1995. *Longman Dictionary of Contemporary English*. Essex: Longman.



- Taljad, E. & G.-M. de Schryver. 2002. Semi-automatic term extraction for the African languages, with special reference to Northern Sotho. *Lexikos* 12:44-74.
- Villavicencio, A., A. Copestake, B. Waldron & F. Lambeau. 2004. Lexical Encoding of MWEs. *Second ACL Workshop on Multi-word Expressions: Integrating Processing*, Barcelona, Spain.80-87.

## CHAPTER 14

# DICTIONARY BASIS AND LEMMATISATION FOR AN ENCYCLOPEDIA OF YILUMBU

Hugues Steve Ndinga-Koumba-Binza<sup>1</sup> and Gilles Saphou-Bivigat<sup>2</sup>

<sup>1</sup>*Centre for Text Technology, North-West University, Potchefstroom, South Africa*

<sup>1</sup>*Langue, Culture et Cognition (LCC), Université Omar Bongo, Libreville, Gabon*  
22602569@nwu.ac.za

<sup>2</sup>*Institut de Recherches en Sciences Humaines, CENARST, Libreville, Gabon*  
saphoubivigat@yahoo.fr

### 1. INTRODUCTION

This chapter focuses on the dictionary basis and a few lemmatisation issues of a planned encyclopaedic dictionary for Yilumbu, a developing language spoken in Gabon and in the Congo. An early study on the planning of an encyclopedic dictionary for this language has been done (cf. Saphou-Bivigat 2002 & 2010). A review of the suggestions contained in the abovementioned study reveals a number of discrepancies on the proposed dictionary basis and the lemma selection. This chapter intends to re-examine the same topics and suggests new perspectives. The dictionary basis is the set of sources “used for the compilation of a dictionary” (Svensén 2009:39). These include primary sources, secondary sources as well as tertiary sources. The aim of the contribution is to indicate which sources should be relevant for the planned encyclopedic dictionary and how items gathered from these sources should be lemmatised. As for lemmatisation, it is referred to in Svensén (2009:5) as the procedure that ensures that the “*selection of words for the dictionary is as representative as possible*” in using the “*total frequency of all the forms of a certain word*”. This definition is expanded in the *Collins English Dictionary*, as the process of grouping together the different inflected forms of a word so that they can be analysed as a single item, i.e. as a lemma in a dictionary.

### 2. DICTIONARY BASIS ISSUES

We have mentioned above that the set of sources constitutes the dictionary basis. Sources refer to the collection of material and they may be oral or written. Encyclopedic dictionaries are compiled from a variety of sources. When dealing with the type of the planned encyclopedic dictionary, lexicographers must be aware of an important fact, i.e. the tradition of language use of the community that speaks the language.

The main question would be: is it a written language or not? The answer to this question would determine the nature of dictionary sources, i.e. what would be regarded as primary sources, secondary sources or tertiary sources. A similar question to this one has been previously dealt with in Nzang-Bie (2002) with regard to lexicographic corpus development in Gabonese languages.

## 2.1. Primary Sources

The nature of primary sources of a dictionary could depend on the language for which the dictionary is made. This would depend on whether the language had enough written resources for the primary sources to be solely written sources. Otherwise, the primary sources would be oral sources. It is, for instance, noticeable that major languages such as English and French have essentially written primary sources for any type of dictionary planned in these languages. Languages with an oral tradition require a different approach from languages with a long writing tradition.

Yilumbu is a language with an oral tradition. Proposals for an alphabet and an orthographic system have been made only for the last two decades (Emejulu & Pambou-Loueya, 1990; Mavoungou, 2002a, 2010a, 2010b & 2010c). Thus, primary sources for any dictionary of Yilumbu would be as indicated in Prinsloo (2000:4), i.e. the recording of a large variety of spoken speech from as many different genre/topic areas as possible. For the planned Yilumbu encyclopedic dictionary and within the framework of his study, Saphou-Bivigat (2010) stated that he successively conducted two fieldwork sessions between December 2001 and January 2003 in the Yilumbu-speaking areas in Gabon. The author selected five native speakers of various ages and educational levels. He maintained a balance between young and relatively old speakers in his study. This was done to ensure an accurate representation of certain important historical aspects of the culture (culture-specific concepts) being described and to gain insight into the present-day reality of the language and living conditions.

## 2.2. Secondary Sources

Secondary sources are written materials and they basically refer to all the dictionaries consulted during the compilation phase (cf. Wiegand & Kučera 1981:100ff). It should be mentioned that these are very important dictionary sources. Whenever the secondary sources exist, they should never be left out of consideration, especially when compiling a dictionary for a minority language such as Yilumbu. That is why, in order to emphasise the importance of secondary sources, Zgusta (1971:239) writes that *"sometimes, one dictionary is the basis for the compilation of another"*.

For Yilumbu, however, not many dictionaries exist. The sole existing Yilumbu dictionary is that of Mavoungou and Plumel (2010). This monoscopal (Yilumbu-French) dictionary has a central list of 6000 lemmas and indexes. This dictionary is a valuable source. In fact, the lexicon recorded in the list integrates both the terms of the basic vocabulary covering everyday life, and those of specific domains (e.g. hunting, fishing, picking) as well as a certain number of loanwords (adopted from Portuguese, English and French in particular). Furthermore, the lexicographical treatment of the various entries creates here and there a set of ethnographical data on Balumbu (the ethnolinguistic group that speaks Yilumbu) and other peoples of Gabon, their ecosystem, local French, etc.

Another important source would be the *Yilumbu Frequency List* compiled by Mavoungou (2000). This list is a structured electronic corpus comprising 35 660 running words (Yilumbu-French). The list, aimed at identifying the most frequent words of the language, was collected in January 2000 as a preliminary survey of the Yilumbu corpus for the work by Mavoungou (2002a). A currently planned idiomatic dictionary of Yilumbu (cf. Mavoungou, in this volume,) is also a source to consider in the planning of a Yilumbu encyclopedic dictionary.

Finally, as suggested in Zgusta (1971:239), the lexicographer comparing his or her own material with other dictionaries should have an attitude of scientific criticism: *"Nothing is to be accepted from another source without a constant checking up of every detail"* (Zgusta 1971:239). In this line of argumentation, it can also be mentioned that a lexicographer may also use a good monolingual dictionary as a major source for the dictionary basis. Here again, and on account of Zgusta's remark, if the lexicographer uses such a monolingual dictionary as a primary source for the encyclopedic dictionary, the selection of lemmata that is presented in the monolingual dictionary should be approached with an attitude of scientific criticism. However, there is no existing Yilumbu monolingual dictionary. And no such dictionary is currently planned in the emerging Yilumbu lexicography. The planning of an encyclopedic dictionary of Yilumbu is therefore confined to bilingual reference works as secondary sources.

### 2.3. Tertiary Sources

Tertiary sources encompass all linguistic monographs, papers and grammars used for the constitution of the dictionary basis (cf. Gouws 2001:69). With regard to tertiary sources, most existing works in Yilumbu deal with religious, pedagogical and scientific literature. The literature on religion includes the work done by Garnier (1897, 1900 & 1904) and Murard (1903a & 1993b). Garnier is the author of three books based on the dialect that is spoken in the Nyanga province (in Mayumba particularly):

- (i) *Katesisa igheghe nesi malonghi ma nzambi mu mbembu i-lumbu* (1897);
- (ii) *Syllabaire i-lumbu keti miganda mio mi teti mi uranganga mu mbembu i-lumbu* (1900); and
- (iii) *M'ambu ma nzambi mo make mukatesisa* (1904).

Contrary to Garnier's work, Murard's books are based on the dialect spoken in the Ogooué Maritime province where one also finds Yilumbu (cf. Mavoungou, 2002a: 139):

- (i) *Katsisu ikeki irendulu mu mbembu bis' Setté-Cama* (1903a) and
- (ii) *Katsisu neni irendulu mu mbembu bis' Setté-Cama* (1903b).

Despite these religious literary works of Garnier and Murard early in the 20<sup>th</sup> century, no Bible has ever been written for Yilumbu. For scientific literature, pioneering works in the field of linguistics date from Blanchon (1984), and Emejulu and Pambo-Loueya (1990). There is, however, no existing comprehensive grammar published of the language. Recent linguistic works are that of Mavoungou (2002b, 2005a, 2010b & 2010c), Saphou-Bivigat (2000) and Mboumba (2009). In spite of the limited number of linguistic works, Yilumbu has benefited from a number of works in the field of lexicography within the metalexicographic planning of a trilingual dictionary of Yilumbu (cf. Mavoungou 2002a, 2002b, 2002c, 2006 & 2010a). Saphou-Bivigat (2002 & 2010) has also recently contributed to the Yilumbu lexicographic literature. Again, no data from any scientific source should be accepted uncritically by a lexicographer.

### 3. LEMMATISATION ISSUES

A few lemmatisation issues for the planned encyclopedic dictionary of Yilumbu are dealt with in this chapter. These include the alphabet issue, the orthography issue, the structure of the lemma and the lemma candidate list.

#### 3.1. Alphabet and Orthography Issues

Deciding about an alphabet and an orthographic system for Yilumbu has been an issue like it is for most of the Gabonese languages. As mentioned earlier, a number of proposals have been made for the Yilumbu alphabet and orthography. Although, none of the proposals has yet received government accreditation to be taught and used in schools, the current proposals do not have an agreed written form for a number of items. For instance, the representation of vowel length is differently suggested by Emejulu and Pambo-Loueya (1990) and Mavoungou (2010c): the former do not suggest any representation of vowel length in the orthography of Yilumbu; the latter suggests doubling the vowel as representation of vowel duration in the language orthography.

The publication of the Yilumbu dictionary by Mavoungou and Plumel (2010) appears as a way to impose the alphabet and the orthographic system contained in it. This dictionary also contains the most recent proposals for the Yilumbu alphabet and orthography. The present work advocates the adoption of these proposals for the planned encyclopedic dictionary of Yilumbu until improvements have been suggested. For instance, Ndinga-Koumba-Binza and Roux (2009) have indicated that any proposal for an alphabet and/or orthography of minority languages should be subjected to experimentation with the speaking communities in the process of its design. A few elements that the authors have suggested as community-acceptable principles for a writing system for Gabonese languages comprise the following (Ndinga-Koumba-Binza & Roux 2009:101):

- (i) Phonological characteristics,
- (ii) An alphabet for orthographic purposes, including writing rules, reading rules, punctuation rules, and capital letters rules,
- (iii) User-friendliness in learning and writing,
- (iv) Minimum of problems in readability,
- (v) Machine printability,
- (vi) Uniformity for all Gabonese languages.

The planned encyclopedic dictionary for Yilumbu would certainly meet a friendliness requirement if the orthography proposed for the language is conceived in terms of the model of the dominant language of the region, in this case French, where the smaller language is spoken. This is a principle suggested by Coulmas (1996) with regard to establishing a writing system for languages of oral tradition. The following randomly chosen article is a typical illustration of the alphabet list and writing system to be used for Yilumbu:

**DIFUGHU DI WIISI WEENDA** [dífúyù dì wí:sì wé:ndà] (Aussi **difughu di wiisi**). (pl. **mafugha ma wiisi weenda**) ■ Le dartrier (*Cassia Alata*). Plante arborescente à feuillage ornemental. **ENC.:** *Difughu di wiisi weenda* signifie littéralement "action de se replier à la tombée de la nuit". Cette belle plante arborescente s'ouvre le matin et se referme le soir. **USAGES:** En décoction, les feuilles du dartrier sont utilisées en lavement et comme un vomitif. Les mêmes feuilles entrent aussi dans le traitement de la dartre, des mycoses et autres maladies de la peau.

One would note in the article above the user-friendly orthographic representation of the Yilumbu phonemes and word division.

### 3.3. Structure of the Lemma: Stem vs. Word

With regard to African languages, two lexicographic traditions exist in lemmatisation, namely the word tradition and the stem tradition. According to the word tradition, lexical items are entered in their complete forms. In other words, prefix plus stem, while in the stem tradition, lexical items are lemmatised under the stem without their prefixes. The figure below illustrates the stem tradition in Pove, a language spoken in Gabon, as applied in Mickala Manfoumbi (2004:44). Column 1 shows the lemmatised stem, column 2 gives an indication of the noun prefix number, column 3 shows the entire word and column 4 gives the French equivalent and/or the encyclopedic information (see Mavoungou 2005b and Ndinga-Koumba-Binza 2006 for detailed reviews on Mickala Manfoumbi's dictionary).

**Figure 1:** Stem tradition in Mickala Manfoumbi (2004:44)

<b>b</b>			
+ bá -		báká	<i>être</i>
+ bà -		báka	<i>épouser</i>
+ bà -		baáka	<i>sarcler</i>
+ bà	7,8	geba biba	<i>endroit calme et profond d'un fleuve</i>
-bà ~ - bálé ~ - bàdí (variante apocopée de - bálé)		bíba (classe 8)	<i>deux</i>
		nzima díba	<i>vingt (deux fois dix)</i>
+ bá - án -		báná	<i>épouser</i>
+ bàbà	3,4	mubaba mibaba	<i>grande lèvre d'un sexe féminin</i>
+ bábà	5,6	ebába mabába	<i>gifle</i>
+ bàbè	7,8	gebabe bibabe	<i>écorce</i>
+ bábàkà	5,6	ebábaka mabábaka	<i>pancréas</i>
+ bábáká	5,6	ebábáká mabábáká	<i>petit manioc (plié en deux)</i>
+ bàb - *úd -		babwáka	<i>débiter, décapiter</i>
+ bàd -		badáka	<i>se couvrir, se protéger</i>

Dictionaries using the stem tradition are usually considered by Gabonese linguists (who mostly do not have any background in metalexicography) as being more scientific than word dictionaries. However, the choice of the stem tradition over the word tradition may lead to access difficulties on the part of users.

A close look at the survey of lexicographic activities in Gabon (cf. Nyangone Assam & Mavoungou 2000; Mihindou 2001; Ndinga-Koumba-Binza 2005, 2006 and 2010; Mavoungou 2010b) shows that lexicographers are more in favour of word lemmatisation than stem lemmatisation. It is herein agreed with Mavoungou (2002a & 2010b) who has emphasised that adopting the stem tradition for the Gabonese languages will have confusing results for users not familiar with stem dictionaries. That is the reason why we adhere to the word tradition for the planned encyclopedic dictionary of Yilumbu. The majority of existing dictionaries in Gabonese languages have used the word tradition for lemmatisation.

It will be part of the editorial policy of the proposed encyclopedic dictionary to lemmatise all the lexical items according to the word tradition.

### **3.3. The Lemma Candidate List**

The lemma candidate list may be referred to as the list of lexical items that are selected by the lexicographer for inclusion as primary treatment units in a given dictionary. The selection of core vocabulary items may rely on the available frequency lists of a given language. Relying on this frequency list by Mavoungou (2000), we will consider the issue of what sort of lexical items should be included in the planned encyclopedic dictionary. The 80 most frequent types of the Yilumbu corpus (extracted from Mavoungou 2000):

<b>Lexical items</b>	<b>Frequencies</b>
na	1 944
mí	1 664
ti	1 111
nge	973
o	970
ya	937
ke	827
mo	745



vho	664
mutu	529
nana	515
mwaana	495
a	481
ka	430
kantsi	383
vhana	383
ti	381
mu	367
ika	359
bwaala	311
yaandi	297
batu	296
ghuna	292
baana	289
aghu	288
o	285
yetu	284
ibili	277
gho	275
ana	267
be	266
maama	264
mwaana	243
i	237
yina	237
lyongu	233
mwa	232

sa	230
maamba	210
vhana	209
vho	204
ma	198
di	192
noongu	192
kala	190
ifuumba	184
mo	184
nguyi	183
pi	181
utuba	179
dyaambu	178
a	173
vhavha	173
mweegha	166
i	165
me	160
dibaala	157
mu	153
yenu	153
yo	150
má	149
ibaamba	144
iboonga	142
mbaatsi	141
mbe	141
uyi	139

minu	130
yaayi	130
vandi	129
ba	126
beeni	126
maambu	123
mughetu	122
mbeembu	120
dí	119
murú	117
musiru	117
fu	116
pweela	115
vangi	115

Merely by going through the 80 most frequent types, one already covers one-third of the entire electronic Yilumbu corpus. An examination of the above list shows that apart from the lexical items *mi* (short for *minu* 'I') and *nge* (short for *ngeeyu* 'you') the top 10 items of the list are purely grammatical words, which as such do not meet the main criterion (of being cultural and extralinguistic items) for inclusion in an encyclopedic dictionary.

Thus, given the nature (encyclopedic) and the genuine purpose (cognitive) of the planned dictionary, it will be unproductive to base the compilation of the lemma sign list only on word frequency counts. Cultural lexical items are indeed the most important for any planned encyclopedic dictionary. In this regard, the lexicographer could rightfully include additional lexical items that do not appear in the frequency counts but are part of the cultural knowledge of the linguistic community, i.e. the users. These lexical items comprise for instance words of specialised jargons in various domains such as fishery, hunting and farming. In the domain of fishery, we have the following examples which indicate the names of different types of sardines.

- (i) dikwaala "*flat sardine*"
- (ii) diseengi "*smallest sardine*"
- (iii) disuungi "*long sardine*"

(iv) dubali "regular sardine"

(v) dyeenga "red sardine"

#### 4. CONCLUSION

This chapter has dealt with the dictionary basis and a number of lemmatisation issues in the planning of an encyclopedic dictionary for Yilumbu. Firstly indicated primary sources, secondary sources and tertiary sources of the planned dictionary were indicated. Relevant sources may be regarded as the ones that will be representative of the linguistic and cultural reality that the target users of the planned dictionary will have to face on a daily basis. As for lemmatisation, numerous suggestions were made for the standardisation of the writing system, the structure of the lemma and the lemma candidate list.

It is herein acknowledged that the planning for a Yilumbu encyclopedic dictionary is sufficiently advanced, but many other issues still have to be treated prior to the final work on the planned dictionary. In fact, as Kavanagh (2002:275) puts it: "*good planning, sustained and systematic editing and plenty common sense go a long way to ensuring a high quality text*".

#### REFERENCES

- Blanchon, J.A. 1984. Présentation du yi-lumbu dans ses rapports avec le yi-punu et le ci-vili à travers un conte traditionnel. *Pholia* 1:7-35. Reprinted in Blanchon, J.A. 1999. *Douze études sur les langues du Gabon et du Congo-Brazzaville*. München: Lincom Europa.5-31.
- Collins English Dictionary*. 2010. 10th Edition. New York: CollinsHarper.
- Emejulu, J. & F. Pambo-Loueya. 1990. Yilumbu. *Revue Gabonaise des Sciences de L'Homme* 2:197-201.
- Garnier, A. 1897. *Katesisa i gheghe nesi ma loghi ma dzambi mu mbembu i-lumbu*. Loango: Imprimerie de la Mission.
- Garnier, A. 1900. *Syllabaire i-lumbu keti mi ganda mio mi teti mi ranganga mu mbembo i lumbu*. Loango: Imprimerie de la Mission.
- Garnier, A. 1904. *M'ambu ma nzambi mo make mu katesisa*. Loango: Imprimerie de la Mission.
- Gouws, R.H. 2001. Lexicographic training: Approaches and topics. *Eléments de lexicographie gabonaise.*, edited by J.D. Emejulu. Tome I. New York: Jimacs-Hillman Publishers.58-95.

- Kavanagh, K.P. 2002. Adapting a monolingual dictionary for local use. In: *Éléments de lexicographie gabonaise*, Tome II, edited by J.D. Emejulu. New York: Jimacs-Hillman Publishers.263-275.
- Mavougou, P.A. 2000. *A frequency list of the Yilumbu language*. Unpublished manuscript. Stellenbosch: Bureau of the Woordboek van die Afrikaanse Taal.
- Mavougou, P.A. 2002a. *Metalexicographical criteria for the compilation of a trilingual dictionary: Yilumbu-English-French*. Unpublished DLitt dissertation. Stellenbosch: Stellenbosch University.
- Mavougou, P.A. 2002b. Sociolinguistic and linguistic aspects of borrowing in Yilumbu. *South African Journal of African Languages* 22(1):41-58.
- Mavougou, P.A. 2002c. Synopsis Articles in the Planning of a Trilingual Dictionary: Yilumbu-English-French. *Lexikos* 12:181-200.
- Mavougou, P.A. 2005a. Loanwords in Yilumbu: A morphological, semantic and lexicographic perspective. *South African Journal of African Languages* 25(4):258-272.
- Mavougou, P.A. 2005b. R. Mickala Manfoumbi: Lexique pove-français/français-pove. *Journal of Education* 4(1):79-86.
- Mavougou, P.A. 2006. A trilingual dictionary Yilumbu-French-English: An ongoing project. *Lexikos* 16:121-144.
- Mavougou, P.A. 2010a. *A dictionary plan for Yilumbu*. VDM Verlag.
- Mavougou, P.A. 2010b. Orthographe, standardisation et confection des dictionnaires en Yilumbu, Yipunu et Civili. In: *Écriture et Standardisation des Langues Gabonaises*, edited by J. Hubert & P.A. Mavougou. Stellenbosch: SUN Press.97-134.
- Mavougou, P.A. 2010c. Propositions pour l'orthographe du Yilumbu. In: *Écriture et Standardisation des Langues Gabonaises*, edited by J. Hubert & P.A. Mavougou. Stellenbosch: SUN Press.182-187.
- Mavougou, P.A. & B. Plumel. 2010. *Dictionnaire yilumbu-français*. Libreville: Editions Raponda-Walker.
- Mboumba, L.H. 2009. *Quelques Problèmes Métalexicographiques en yilumbu: le cas des parties du discours*. Unpublished MA thesis. Libreville: Université Omar Bongo.
- Mickala Manfoumbi, R. 2004. *Lexique pove-français/français-pove*. Libreville: Editions Raponda-Walker.

- Mihindou, G.-R. 2001. Apports des missionnaires à la lexicographie gabonaise: Dictionnaires bilingues fang-français / français-fang; français-yipounou / yipounou-français; français / mpongwe. In: *Éléments de lexicographie gabonaise*, Tome I, edited by J.D. Emejulu. New York: Jimacs-Hillman Publishers.7-37.
- Murard, P. 1903a. *Katsisu i keki i rendilu mu mbembo bis'Sette-Cama* (petit catéchisme). Lyon: Imprimerie Paquet.
- Murard, P. 1903b. *Katsisu i neni i rendilu mu mbembo bis'Sette-Cama* (grand catéchisme). Lyon: Imprimerie Paquet.
- Ndinga-Koumba-Binza, H.S. 2005. Considering a lexicographic plan for Gabon within the Gabonese language landscape. *Lexikos* 15:132-150.
- Ndinga-Koumba-Binza, H.S. 2006. Lexique pove-français/français-pove. Mickala Manfoumbi: Seconde note de lecture. *Lexikos* 16:293-308.
- Ndinga-Koumba-Binza, H.S. 2010. Trends in Gabonese modern lexicography. Paper presented at the 15<sup>th</sup> International Conference of the Association for African Lexicography (AFRILEX) held at the University of Botswana, Gaborone, Botswana: 19-21 July 2010.
- Ndinga-Koumba-Binza, H.S. & J.C. Roux. 2009. On Writing Gabonese languages. In: *Language, Literature and Society*. Proceedings of the First International Conference of the Department of African Languages and Literature, University of Botswana, Gaborone, Botswana, 26-28 June 2008, edited by H. Batibo, R. Dikole, S. Lukusa & Nhlekisana. Gaborone: Associated Printers.83-106.
- Nyangone Assam, B. & P.A. Mavoungou. 2000. Lexicography in Gabon: A survey. *Lexikos* 10:252-274.
- Nzang-Bie, Y. 2002. Le corpus lexicographique dans les langues à tradition orale: le cas du dialecte fang-mekè. *Lexikos* 12:211-226.
- Prinsloo, D.J. 2004. *The compilation of electronic corpora, with special reference to the African languages*. Class notes for the lexicography module. Department of Afrikaans and Dutch. University of Pretoria/Stellenbosch University.
- Saphou-Bivigat, G. 2000. *L'emprunt lexical du lumbu, langue bantoue du Gabon (B44) au français*. Unpublished MA thesis. Libreville: Université Omar Bongo.
- Saphou-Bivigat, G. 2002. Le dictionnaire encyclopédique: Théorie et modèle dans les langues gabonaises. In: *Éléments de lexicographie gabonaise*, Tome II, edited by J.D. Emejulu. New York: Jimacs-Hillman Publishers.174-186.

- Saphou-Bivigat, G. 2010. *A theoretical model for an encyclopaedic dictionary for the Gabonese languages with reference to Yilumbu*. Unpublished DLitt dissertation. Stellenbosch: Stellenbosch University.
- Svensén, B. 2009. *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge, UK: Cambridge University Press.
- Wiegand, H.E. & K. Kučera. 1981. Brockhaus-Wahrig: Deutsches Wörterbuch auf dem Prüfstand der praktischen Lexikologie. II. Teil: 1 Band (A-BT): 2. Band (BV-FZ). *Kopenhagener Beiträge zur Germanistischen Linguistik* 18:94-217.
- Zgusta, L. 1971. *Manual of lexicography*. The Hague: Mouton.

## CHAPTER 15

# INNOVATIVE STRATEGIES IN MACROSTRUCTURAL CHOICES

Rufus H. Gouws

*Department of Afrikaans & Dutch, Stellenbosch University, Stellenbosch, South Africa*  
rhg@sun.ac.za

### 1. INTRODUCTION

When planning and compiling a dictionary a lexicographer can follow either a contemplative or a transformative approach, cf. Tarp (2000). Where the first choice restricts the lexicographer to previously established ways and means of selecting data and presenting it in the envisaged dictionary the latter offers the opportunity to be innovative and, without denying the value of certain aspects of established lexicographic work, introduce features that can enhance the quality of the envisaged dictionary as a practical instrument, directed at a specific target user group. The intended target user remains the dominating factor when planning and compiling a dictionary. In accordance with the needs and reference skills of this target user the lexicographer decides on the function of the dictionary. In order to satisfy the identified function the data to be included in the dictionary is selected and a decision is made regarding the data distribution. This is followed by a selection of the structures needed to accommodate and present the data and to ensure the best possible access to the data. When user needs, functions, data and structures are negotiated in an appropriate way a dictionary can be compiled that will see the users achieving an optimal retrieval of information and successful dictionary consultation procedures.

In printed dictionaries the macrostructure is one of the primary structures. However, this structure is too often regarded as a mere ordering of lemmata, without the necessary attention to the different possibilities that prevail on macrostructural level. The advent of e-lexicography has had a definite influence on the interest in macrostructural issues. One of the many differences between online dictionaries and their printed counterparts resorts in the lexicographic structures. Whilst some of the structures of printed dictionaries are also relevant to online dictionaries, albeit in an adapted way, one has to recognise the fact that some other structures play a different role in online dictionaries compared to printed dictionaries. Online dictionaries, and the CD ROM versions of printed dictionaries are excluded from this remark, do not have a macrostructure. The lack of a macrostructure in online dictionaries has decreased the interest in having the macrostructure as a topic in theoretical discussions. However, this limited interest



has not withheld practical lexicographers from employing new macrostructural strategies in their dictionaries. In lexicography theory has often followed practice and although this situation has changed to some extent with theory often playing a leading role and showing practical lexicographers new ways and means to compile their dictionaries, theory can still benefit from taking cognisance from practice, also with regard to macrostructural issues.

This contribution focuses on a number of macrostructural issues, with specific reference to the macrostructural approach in the recently published latest edition of the *Woordeboek vir die Gesondheidswetenskappe/Dictionary for the Health Sciences*.

## 2. THE CURRENT SITUATION

Looking at the way in which the majority of current-day dictionaries have employed their macrostructures it offers little reason for enthusiastic lexicographic or research endeavours. Even the traditional choice between an alphabetic and a thematic ordering of the macrostructural elements no longer plays a major role in current lexicographic practice dominated by an alphabetical approach.

A choice in favour of alphabetical ordering demands a further decision. This concerns the issues of a strict initial alphabetical, a straight alphabetical ordering and subsequently the use of a niched or a nested sinuous lemma file. When opting for procedures of niching or nesting in order to produce a niche- or nest-alphabetical macrostructure, the lexicographer needs to make provision for either grouped or non-grouped nested or niched articles, leading to condensed or non-condensed niche- or nest-alphabetical macrostructures respectively, cf. Bergenholtz, Tarp and Wiegand (1999).

In theoretical discussions some of the most salient contributions in this regard are Wiegand (1989, 1998, 2010) and Bergenholtz *et al.* (1999) and Wiegand and Gouws (2011; 2013). Wiegand (1989:383-393) presents different types of macrostructures. Bergenholtz *et al.* (1999) take this classification further and discuss it in more detail. Wiegand and Gouws (2011) identify further types of macrostructures following an investigation of numerous dictionaries. Some of the types discussed in this classification will be mentioned in the present paper.

Wiegand (1989:372) defines a macrostructure as an ordering structure of which the structure carrying set is a non-empty finite set of carriers of guiding elements of a lexicographic reference work. He allocates alphabetical macrostructures the additional general genuine function of showing the lexicographic coverage of the dictionary. This is ensured by the appropriate selection of items to be included in the macrostructure.

### **3. LEMMAS REPRESENTING DIFFERENT TYPES OF LEXICAL ITEMS**

The lexicon of a language includes all the lexical items of that language. These items do not all belong to the same type. In a language like Afrikaans a distinction can be made between words, items smaller than words (i.e. stems and affixes) and multiword lexical items (i.e. loanword groups, group prepositions and fixed expressions; and multiword terms in languages for special purposes). In the planning of dictionaries the lemmatisation of words and items smaller than words usually does not pose problems. They are included as lexical and sublexical lemmata respectively. In a language like Afrikaans or English it is usually quite clear whether an item should be included in a dictionary as a lexical or a sublexical lemma. In some cases different variants of a lexical item need to be lemmatised, leading to a single lexical item being included as a lexical and a sublexical lemma, cf. Gouws (1989:88). Within the African languages, lexicographers have to make a well-informed choice when it comes to the issue of word or stem lemmatisation. This topic has been discussed extensively in a number of publications, e.g. Prinsloo (1994), Prinsloo and Gouws (1996), and various suggestions have been made to enhance the process. Whether lexical items are lemmatised as lexical lemmata or sublexical lemmata usually does not have an effect on the nature of the macrostructure because the lemma signs of both lexical and sublexical lemmata can easily be ordered within a straight alphabetical ordering.

When dealing with lexical lemmata, lexicographers need to negotiate the position of complex words with regard to their lemmatisation. Complex words are often included as nested or niched lemmata with the nests or niches attached to the article of a lemma that represents the lexical item occurring as first stem of the complex word. Although these nests and niches are usually attached to the articles of lexical lemmata such a procedure can also prevail with regard to sublexical lemmata, where the item represented by the sublexical lemma functions as first component of the complex word. The extent of the treatment allocated to niched and nested lemmata is often determined by the type of dictionary. In a bilingual dictionary these lemmata receive a limited treatment with a focus on the translation equivalent. In monolingual dictionaries the nested or niched lemmata often represent semantically transparent multiword lexical items and they are entered as so-called self-explanatory items with their very limited treatment not even providing a paraphrase of meaning. In dictionaries dealing with specialised languages the complex terms are formally linked to related terms and treated as members of a category. In spite of this method of presentation, these niched or nested entries remain lemmata and irrespective of being main or sublemmata they are macrostructural elements of the specific dictionary.

In contrast to the lemmatisation of words and items smaller than words, the lemmatisation of multiword lexical items can have a definite influence on their macrostructural position. Different dictionaries have different domestic conventions in this regard. Typically some multiword units, e.g. loanword groups like *nolens volens*, *ex gratia*, *ex post facto*, and in some cases even group prepositions like the Afrikaans items *met betrekking tot*, *na aanleiding van*, *ten spyte van* are entered as lemmata in the alphabetical position where the first member of the multiword unit belongs. This is not the default procedure in all dictionaries. Fixed expressions and here the term is used to include e.g. idioms and proverbs are fully-fledged lexical items and the way they are presented in a dictionary should reflect this status. The typical way in which fixed expressions are entered, cf. Gouws (1989, 1996, 2010), Botha (1991), L. Gouws (2006), Bergenholtz and Gouws (2007) does not reflect their status as fully-fledged lexical items but rather implies that they are presented as part of the lexicographic treatment of a single word occurring in the fixed expression. Including a fixed expression like *to bark up the wrong tree* in the article of the lemma sign *tree*, could be seen as an implication that there is a semantic relation between the fixed expression and the word represented by the lemma sign of the specific article. This way of presenting fixed expressions eschews their status as lexical items.

The lemmatisation of fixed expressions could be done in such a way that their macrostructural status is reflected sufficiently. Botha (1991) suggested such an approach for the WAT, the multivolume comprehensive monolingual dictionary of Afrikaans. Gouws (2010) argues in favour of the inclusion of two separate text blocks of nested lemmata for complex words and fixed expressions, attached to an article of the word functioning as guiding element for the relevant fixed expressions and complex lexical items presented as niched or nested lemmata. This approach would lead to a procedure of comprehensive nesting with the nesting of fixed expressions complementing the nesting of complex words.

Both Botha (1991) and Gouws (2010) represent innovative changes within the default macrostructures of the respective dictionaries to account for a specific type of lexical item. Innovation can go beyond the default and lead to a different distribution of lexical items during the lemmatisation processes and a macrostructural coverage resulting in more than one macrostructure in a given dictionary or the amalgamation of different macrostructures into one single structure. The first of these two approaches comes to the fore in the *Grondslagfasewoordeboek* while both approaches prevail in the *Groot woordenboek Nederlands Afrikaans*. These procedures will not be discussed in this paper.

#### 4. THE LEMMATISATION OF MULTIWORD TERMS

Gouws (to be published) has indicated that in the development of a dictionary culture the scope has too often been restricted to the domain of LSP dictionaries, i.e. dictionaries dealing with languages for general purposes. There is a real need to expand the scope so that LSP dictionaries, i.e. dictionaries dealing with languages for special purposes, should also be included. Innovative macrostructural strategies are also needed in LSP dictionaries.

In their discussion of specialised dictionaries Bergenholtz, Tarp and Wiegand (1999) make provision for alphabetical as well as systematic macrostructures. When discussing alphabetical macrostructures Bergenholtz, Tarp and Wiegand (1999:1817) distinguish between two types of niche-alphabetical macrostructures, i.e. condensed niche-alphabetical and non-condensed niche-alphabetical macrostructures. The first type prevails in word lists with grouped niches whereas the second prevails in word lists with non-grouped niches. LSP dictionaries are compiled for well-defined target user groups who typically possess the necessary dictionary using skills to utilise the specific LSP dictionary to its full potential. The macrostructural strategies in LSP dictionaries are not the typical ones to be found in dictionaries for users not yet familiar with at least intermediate dictionary using skills.

The data distribution in any dictionary needs to be done in a consistent way that allows the predictability principle to come into play once the users' guidelines have been consulted. This consistent application of e.g. macrostructural strategies does not imply the consistent application of only a single approach. In general dictionaries the use of niched and nested lemmata in certain partial article stretches coincides with the use of a straight alphabetical ordering in other partial article stretches. This illustrates the occurrence of more than one macrostructural ordering principle in a single word list.

The ordering of lemmata goes hand in hand with the macrostructural coverage of a given dictionary and the way in which the dictionary accounts for the items from the domain of the dictionary subject matter.

Bergenholtz *et al.* (1999:1817) give an example of the way in which multiword terms are presented as non-grouped niches in Wiesner and Riibeck (1991), a dictionary of veterinary medicine, with the niche attached to the article of the lemma sign representing the first word in the multiword term. Multiword terms like *Ancylostoma braziliense*, *Ancylostoma caninum* and *Ancylostoma duodenale* are lemmatised in a non-grouped niche, i.e. not horizontally ordered within the niche as in a grouped niche, attached to the article of the lemma sign *Ancylostoma*. Albeit an application of a non-condensed niche-alphabetical macrostructure, on account

of the vertical ordering of the niched lemmata within the niche, a degree of textual condensation does prevail because the first word of the multiword term is abbreviated to *A.*, resulting in these niched lemmata being presented as *A. braziliense*, *A. caninum* and *A. duodenale*.

In the remainder of this paper the focus will be on different lemmatisation procedures found in one specific dictionary, i.e. the *Woordeboek vir die Gesondheidswetenskappe/Dictionary for the Health Sciences*. Although the approach will primarily be of a contemplative nature the idea is to describe some of the interesting lemmatisation and macrostructural procedures in this dictionary, with an indication of some of the new lemma types resulting from these applications, in order to provide a theoretical discussion from which future lexicographers can benefit.

## 5. MACROSTRUCTURAL STRATEGIES IN THE *WOORDEBOEK VIR DIE GESONDHEIDSWETENSKAPPE / DICTIONARY FOR THE HEALTH SCIENCES*

### 5.1. Clustering

The recently published second edition of the *Woordeboek vir die Gesondheidswetenskappe/Dictionary for the Health Sciences* (=Brink and Lochner 2011<sup>2</sup>) contains numerous multiword terms. The occurrence of these terms has compelled lexicographers to employ both traditional and innovative lemmatisation strategies. Although this dictionary primarily has a vertically ordered main alphabetical macrostructure the use of a wide-ranging variety of niched and nested articles with niche and nest lemmata as guiding elements, ensures a hybrid macrostructure.

In the *Woordeboek vir die Gesondheidswetenskappe/Dictionary for the Health Sciences* (henceforth abbreviated as *WGW*), multiword lexical items are frequently lemmatised as part of the default ordering within the macrostructure. Terms like *tetralogie van Fallot*, *tiende kraniale senuwee*, *zygomaticus minor* are included as main lemmata in the article stretch where the first component of these multiword items belongs. Due to morpho-semantic and space-saving reasons, multiword terms that share a mutual component are frequently included within non-grouped article clusters presented in either a niched or a nested format. The *WGW* employs non-grouped clusters instead of grouped clusters, because each one of the lemmata participating in these clusters is ordered within the niche or nest at the beginning of a new line and they are not presented in a sinuous lemma file.

These clusters are often attached to the article of the lemma sign representing the first word of the multiword term. As a further space-saving attempt the mutual first word of the multiword terms is not repeated but presented in an abbreviated form, cf. the niche attached to the article of the lemma sign *caput*, i.e. *c. medusae*, *c.*

*obstipum*, *c. succedaneum*. The abbreviated form of the first component transforms the item to a partial lemma, cf. Bergenholtz, Tarp and Wiegand (1999:1816). Where a non-grouped niche normally contains main lemmata, cf. Bergenholtz, Tarp and Wiegand (1999:1815), the partial lemmata *c. medusae*, *c. obstipum*, *c. succedaneum* are sublemmata of the full lemma, i.e. the main lemma *caput*, because access to and interpretation of the sublemma is only possible via the main lemma. A similar procedure is also used in the ordering of complex words being presented as sublemmata in a niche, with a hyphen following the abbreviated first component to indicate that it is part of a compound, presented as sublemma. The article of the simplex *heup* is followed by the niched partial article stretch *h.-bad*, *h.-been*, *h.-gewrig*, *h.-gewrigsiekte*, *h.-jig*, representing the compounds *heupbad*, *heupbeen*, *heupgewrig*, *heupgewrigsiekte* and *heupjig*.

By means of a well-planned deviation of the way in which multiword terms with a mutual first component are presented as niched partial lemmata, the WGW also makes provision for the presentation of some multiword terms with a mutual first component as full lemmata in a partial article stretch attached to the article of a base lemma, cf. Wiegand (1989:390). This procedure is employed where a term being entered in an article cluster is a direct loan from Latin or Greek, whereas the relevant base lemma has been adapted to the Afrikaans orthography or morphology. In these cases the compound with the Latin/Greek version is presented in an unabbreviated form, resulting in a full lemma. As an illustration: attached to the article of the base lemma *makula* (adapted to the Afrikaans spelling) a partial article stretch with the following lemmata, representing compound terms and functioning as guiding elements of these articles, are included: *macula acustica*, *macula albida*, *macula caeruleae*, *macula communicans*, *macula corneae*, *macula densa*, *macula lutea* and *macula retinae*. In all these multiword terms the original Latin spelling prevails. Although these lemmata are all full lemmata they still are sublemmata due to the alphabetical deviation with regard to the preceding base lemma and the subsequent vertically-ordered main lemma, i.e. *makulasie*. In terms of Gouws and Prinsloo (2005:109) this is an application of first level nesting. The WGW follows an innovative macrostructural procedure to ensure that the multiword terms are given in their established form but that they are linked to the orthographically-adapted base lemma. This procedure is carefully explained in the users' guidelines text of the WGW.

Where multiword terms are presented in a niche that is attached to the article of the lemma sign representing the first word of the multiword term, and this first word is an adjective that is declined in its occurrence within the multiword term, the condensed sublemma reflects the declination. Attached to the article of the lemma sign *alkoholies* the following sublemmata, presented as partial lemmata, are guiding elements of a partial article stretch presented in a non-grouped niche-alphabetical ordering: *a. (-e) demensie*, *a. (-e) hallusinose*, *a. (-e) psigose*. The unabbreviated full

form would be *alkoholiese demensie*, etc. The combination of presenting a partial lemma and employing a process of textual condensation by means of a procedure of abbreviations of one or more than one of the lemma parts, demands a much higher level of dictionary using skills from the intended target user, and lexicographers must be sure that their target users have the necessary dictionary using skills. In a dictionary like the WGW the lexicographer may assume that the target users are well-trained users of reference works, including dictionaries. However, it remains important that the different lemmatisation procedures should be discussed in the users' guidelines text. This is largely done in the WGW.

## 5.2. Mutual Proper Nouns

Yet another innovative procedure of lemmatisation is employed in the inclusion of a group of terms that have a mutual word as first component and where this word is a proper noun but where this proper noun in its occurrence as simplex has not been included as a term. WGW includes many terms that are nominals of the type *Gifford se refleks* (Gifford's reflex), *Prussak se vesels* (Prussak's fibres), and quite often the genitive form occurs in a group of terms, although it does not occur outside a multiword combination as term. In such a case the term that comes first in the relevant alphabetical ordering is included as a main lemma and it receives the default treatment. The alphabetically subsequent terms are not lemmatised as main lemmata but as sublemmata, condensed to partial lemmata by means of a procedure of abbreviation of one of the lemma parts, in a non-grouped niche attached to the article of the first member of the group. As an illustration of this procedure: the multiword term *Jacobson se kanaal* (Jacobson's canal) is presented as a main lemma. Attached to the article of this lemma is a non-grouped niche including a partial article stretch with the following condensed sublemmata: *J. se kraakbeen*, *J. se orgaan*, *J. se senuwee* *J. se sulkus*. Each one of the subarticles in this niche contains the relevant default treatment addressed to the sublemma, presented as a partial lemma. This approach of including only one term as main lemma and the others as condensed sublemmata is followed even when there are only two terms in a specific group, e.g. *Jackson se epilepsie* is given as main lemma with a single sublemma, i.e. the partial lemma *J. se sindroom*, and its treatment attached to this article. In terms of Wiegand (1989:388) this sublemma is not part of a niche because a niche contains at least two articles. This type of lemmatisation leads to a single macrostructurally-isolated sublemma condensed to a partial lemma by means of a procedure of abbreviation. The fact that this lemma occurs at the beginning of a line defies the existence of a sinuous lemma file but the fact that it is a sublemma disqualifies a possible straight-alphabetical ordering within this partial article stretch because access to this lemma is only possible via the preceding main lemma. With such an application very little space-saving is achieved but a relation between the two terms is explicitly indicated.

The hybrid nature of the macrostructure of the WGW comes to the fore in different ways. The approach followed in the lemmatisation of the terms mentioned in the previous paragraph is also used where the group of terms (even when there are only two terms in the group and the lemmatisation leads to the inclusion of a single macrostructurally-isolated sublemma) includes a compound term and a multiword term, with the stem variant of a proper noun as first component of the compound and the word variant of that proper noun functioning as the noun in a genitive construction of a multiword term. As an example, the compound term *Hortega-sel* is included as main lemma. The component *Hortega*, occurring as a stem in this compound, also occurs as a proper noun within a genitive in the multiword term *Hortega se metode*. This term is included as a partial lemma and a sublemma, being the guiding element of an article attached to the article of the main lemma *Hortega-sel*. This procedure is also followed where the mutual element, i.e. the first stem of the compound and of the complex proper noun in the genitive form, is a morphological complex form. This applies to the first stem *Morax-Axenfeld-* of the compound *Morax-Axenfeld-konjunktivitis* and the first component *Morax-Axenfeld* of the multiword term *Morax-Axenfeld se diplokokkus*. In accordance with this approach the article of the full lemma *Morax-Axenfeld-konjunktivitis* has a single macrostructurally-isolated sublemma condensed to a partial lemma by means of a procedure of abbreviation attached to it, i.e. the partial lemma *M-A se diplokokkus*. Here the main lemma *Morax-Axenfeld-konjunktivitis* is also the base lemma of the partial lemma *M-A se diplokokkus* through which access to this partial lemma is possible.

This type of macrostructural procedure, i.e. with a complex mutual first component where the word variant of this component is a proper noun, is taken a step further where the mutual element is a complex lexical item that becomes the first component of an extended complex stem, i.e. where a complex stem combines with another stem to be the first stem of a compound, in the sublemma that is condensed to a partial lemma. The compound *Charcot-Marie-simptoom* is included as a main lemma. The complex lemma part *Charcot-Marie-* also functions as part of an extended complex stem in the term *Charcot-Marie-Tooth se progressiewe spieratrofie*. This term is included and presented as a sublemma which is the guiding element of an article attached to that of the main lemma *Charcot-Marie-simptoom*. However, this sublemma is presented as a partial lemma, i.e. *C.-M.-Tooth se progressiewe spieratrofie*, and being the only sublemma in the text block it also constitutes a single macrostructurally-isolated sublemma condensed by a procedure of abbreviation to a partial lemma. Yet again, such a macrostructural procedure demands a user with above average dictionary using skills.

The way in which a group of multiword terms is lemmatised by presenting the term that fills the first alphabetical spot as a main lemma, e.g. *Jacobson se kanaal*, and the following members of the group as non-grouped sublemmata, e.g. *J. se*



*kraakbeen*, *J. se orgaan*, *J. se senuwee*, *J. se sulkus*, is not applied in all occurrences of groups of multiword terms with a similar first component. In the group *Frey se gastriese follikels*, *Frey se hare*, *Frey se sindroom* all three terms have a similar first word, i.e. *Frey*. In terms of the procedure followed in the lemmatisation of *Jacobson se kanaal*, *J. se kraakbeen*, *J. se orgaan*, *J. se senuwee*, *J. se sulkus* one would have expected the occurrence of *Frey se gastriese follikels* as base lemma with *Frey se hare* and *Frey se sindroom* presented as niched lemmata. This has not happened here because each one of these multiword terms is presented as a main lemma. A closer look at the treatment given to these main lemmata indicates that in each article reference is made to a different person called *Frey*, i.e. H. Frey, M. von Frey and L. Frey. A difference in the reference value of the proper nouns functioning within the genitive compels the lexicographer not to create a macrostructural link between them by entering the second and third as sublemmata attached to the article of the first as main lemma. The lack of a semantic link between these terms defies their lemmatisation within a non-grouped article niche. Consequently they are rather included as separate main lemmata.

### 5.3. Lemmatising According to a Non-Initial Lemma Part

#### 5.3.1. First level nesting

Deviation from a method of exclusive straight-alphabetical ordering in the WGW also leads to other types of partial lemmata than discussed up to now, and also to the use of procedures of nesting within the article clusters. The nesting can be applications of first level nesting but it can also be applications of second level nesting, cf. Prinsloo and Gouws (2005:109).

First level nesting results in a deviation of alphabetical ordering between the last nested lemma and the following lemma in the default vertical ordering. Within the nest the alphabetical ordering is maintained. A typical application of second level nesting usually also sees a deviation from the alphabetical ordering between the last nested lemma and the next lemma in the default vertical ordering, but the distinctive feature of second level nesting is a deviation from alphabetical ordering within the given cluster of articles. In WGW first level nesting is much more frequently used than second level nesting. An example of second level nesting can be found in the partial article stretch attached to the article of the lemma *endositose*. This will not be discussed here.

As seen in the previous section, article clusters containing multiword terms and compound terms are often attached to the article of a single word which also functions as first word of the multiword term or of which the stem variant functions as first component of the compound. In the niched lemmata of the WGW this mutual component is often presented in an abbreviated form with the partial

lemma presenting the unique component of the term in full. Many multiword and compound terms have their semantic core in the second component of the term, and a frequent way of lemmatising compound and multiword terms is by attaching their articles to the article of the lemma that represents the simplex form of which a bound variant forms the second component of the compound or multiword term. This leads to a type of first level nesting where a nest-internal alphabetical ordering is maintained but the nest deviates from alphabetical ordering of the partial article stretch into which it fits. The term *kodering* is lemmatised in its alphabetical position, with *kodon* presented as the next main lemma. Attached to the article of the lemma *kodering* is an article nest which contains lemma signs to represent the compound terms *plekkodering* and *skynkodering*. These terms are lemmatised as follows: *k.*, *plek-:* and *k.*, *skyn-:*; with *k* being the abbreviated form of the lemma part *kodering* and the hyphen attached to *plek* and *skyn* functioning as place-keeping symbol for the abbreviated preceding lemma part. These are highly condensed partial lemmata but the system is explained in the users' guidelines text and this form of lemmatisation is done in a systematic way.

Within the article cluster the entries *k.*, *plek-:* and *k.*, *skyn-:* adhere to alphabetical ordering but this cluster does not fit into alphabetical ordering of the surrounding partial article stretch. In articles like these the WGW presents non-grouped partial lemmata that are ordered in terms of the alphabetical value of a non-initial part. It is important to refer to "a non-initial" rather than "a second" or "a final" component. The position of the part in the term according to which the term is alphabetised cannot always be given unambiguously as, say, second or last. Due to this varying position of the form according to which alphabetical ordering is done, this type of lemma is indicated as being alphabetised in terms of a non-initial part.

The use of non-grouped partial lemmata that are ordered in terms of the alphabetical value of a non-initial part occurs frequently in the lemmatisation of multiword terms. Attached to the article of the lemma *gastritis* an article nest includes the following sublemmata condensed to partial lemmata by means of a procedure of abbreviation: *g.*, *akute*; *g.*, *chroniese* and *g.*, *flegmoniese* to represent the multiword terms *akute gastritis*, *chroniese gastritis* and *flegmoniese gastritis*. The lemmatisation of multiword terms displays varying degrees of textual condensation. The term *Coombs se toets* is presented as a multiword main lemma. Attached to its article a nest is given to account for the terms *direkte Coombs se toets* and *indirekte Coombs se toets*. These terms are lemmatised as the partial lemmata *C.s.t.*, *direkte* and *C.s.t.*, *indirekte* respectively. The lemmatisation leads to a non-grouped article nest consisting of sublemmata, condensed to partial lemmata by means of a procedure of textual condensation that abbreviates every word in the multiword mutual second component of the term, and alphabetised within the macrostructure in terms of the non-initial component of the term and within the nest in terms of the term-initial unique components.

### 5.3.2. A single word mutual component

In partial lemmata like *C.s.t., direkte* and *C.s.t., indirekte* the mutual component is a multiword form and the unique component consists of a single word. Another category of multiword terms in the WGW has a single word mutual component and a multiword unique component. Where the mutual form prevails in the non-initial position the lemmatisation of these forms also lead to alphabetisation within the macrostructure in terms of the non-initial component but the textual condensation by means of component abbreviation is restricted to that of a single word. Attached to the article of the lemma *senuwee* the sublemma *s., posterior interosseuse* represents the term *posterior interosseuse senuwee*. However, whether the abbreviation is with regard to the single word or the multiword component of a multiword term does not really influence the comprehension of the partial lemma. For both partial lemmata *C.s.t., direkte* and *s., posterior interosseuse* comprehension is dependent on the preceding main lemma. Although a partial lemma with an abbreviated multiword component might look more condensed than a partial lemma with an abbreviated single word component the retrieval of the full form depends in both cases on going back to the main lemma. For the intended target user of this dictionary it should not be an impediment in the access process.

A multiword term like *Jacobson se kanaal* has a structure XY, with X = *Jacobson se* and Y = *kanaal*. Multiword terms belonging to the same group on account of their sharing the mutual component X are lemmatised as partial lemmata according to the alphabetical position of X, cf. the niche containing the lemmata *J. se kraakbeen*, *J. se orgaan*, *J. se senuwee* and *J. se sulkus*. A multiword term like *direkte Coombs se toets* also has the structure XY but with X = *direkte* and Y = *Coombs se toets*. Multiword terms belonging to the same group on account of their sharing of the mutual component Y are lemmatised as partial lemmata according to the alphabetical position of Y, cf. the nest containing the lemmata *C.s.t., direkte* and *C.s.t., indirekte*. A multiword term like *posterior interosseuse senuwee* is lemmatised in terms of the alphabetical position of the nominal component in term-final position: *senuwee*.

### 5.3.3. Discontinuous lemma parts

The sublemmata condensed to partial lemmata in the article clusters in the WGW, irrespective of their being niched or nested clusters, that have been discussed up to now, all present the parts of the non-abbreviated component of the lemma in such a way that this component functions as a whole as lemma part, preceding or following the abbreviated mutual form that functions as lemma part and guiding element to the lemma sign, cf. *opponens digiti minimi pedis* in *m. opponens digiti minimi pedis* (the lemmatisation of the multiword term *musculus opponens digiti*

*minimi pedis*) and *posterior interosseuse* in *s., posterior interosseuse* (the lemmatisation of the multiword term *posterior interosseuse senuwee*). In these lemmata the non-abbreviated component of the multiword term is represented by a continuous presentation which functions as a whole as either a pre- or a postdeterminer of the noun represented by the abbreviated lemma part.

The component *posterior kutane senuwee* of the multiword term *posterior kutane senuwee van die arm* is structurally comparable to the multiword term *posterior interosseuse senuwee*, being an adjectival predeterminer of the noun *senuwee*. In the multiword term *posterior kutane senuwee van die arm*, the component *van die arm* is a prepositional postdeterminer of the noun *senuwee*. The noun *senuwee* is the core element of this nominal phrase. The lemmatisation is done in terms of the alphabetical position of this nominal core, albeit that the noun *senuwee* is neither in the initial nor in the final position of the multiword term, i.e. not in the typical position of parts of the multiword term chosen to be the guiding elements of the multiword partial lemma.

The lemmatisation of the multiword term *posterior kutane senuwee van die arm* displays a new macrostructural procedure:

*s., posterior kutane, van die arm* (attached to the article of the lemma *senuwee*)

A similar procedure is followed in the following examples:

*s., Münchhausen se, deur prokurasie* (attached to the article of the lemma *sindroom*; representing the multiword term *Münchhausen se sindroom deur prokurasie*)

*k., groei-, van 'n populasie* (attached to the article of the lemma *koers*; representing the multiword term *groeikoers van 'n populasie*).

In the examples of multiword terms given above, the part of the term that has determined the alphabetical position of the sublemmata and has been presented as an abbreviated lemma part has been underlined.

In all these lemmata the non-abbreviated parts constitute two discontinuous lemma parts that represent parts of the term to function as pre- and postdeterminers of the noun that is the semantic core of the multiword term. In these examples the WGW offers a new type of lemma, i.e. a non-grouped nested sublemma condensed to a partial lemma by a procedure of abbreviation, ordered according to a non-initial lemma part which is given in abbreviated form and is complemented by two lemma parts that represent discontinuous parts of the multiword term.

These lemmata challenge the dictionary user and demand sophisticated dictionary using skills. In a lemma like *s., posterior kutane, van die arm* the comma between the lemma parts *posterior kutane* and *van die arm* does give the user an indication that the subsequent lemma parts are in a discontinuous relation. The WGW is not consistent in the use of this comma, as can be seen in the following examples:

*s., dorsale digitale van die voet* (= *dorsale digitale senuwee van die voet*; attached to the article of *senuwee*)

*s., mediale kutane van die voorarm* (= *mediale kutane senuwee van die voorarm*; attached to the article of *senuwee*).

These parts of the multiword terms that determine the alphabetical position of the lemma are in different positions of the multiword term – not all in the second or last position. Consequently this lemmatisation procedure is described as a procedure where the position is determined by a non-initial part – and not a second or last part.

#### 5.4. Mixed Macrostructural Procedures in a Single Non-Grouped Partial Article Stretch

In the preceding paragraphs it has been shown that the WGW utilises procedures of textual condensation by means of abbreviation to create abbreviated lemma parts. Irrespective of the position of these parts in the multiword terms they are presented in the entrance to the partial lemma because the ordering of these sublemmata is done according to either its initial part (e.g. *caput medusae* presented as *c. medusae*) or one of its non-initial parts (e.g. *plekkodering* presented as *k., plek-*).

The WGW sometimes employs both these types of lemmatisation in a single cluster, leading to a first level nest with lemmata ordered within the nest according to a mutual lemma part, but where that mutual lemma part does not have the same position in all the nested lemmata. An article nest attached to the article of the lemma sign *foramen* contains, among others, the following sublemmata:

*f., anterior kondilêre*

*f., apikale*

*f. caecum*

*f. incisvium*

*f., jugulêre*

The ordering within the nest is not determined by the nature of the lemma but only by the alphabetical value of the unique lemma part. The user has to interpret the use of only a full stop after the abbreviated lemma part or a full stop as well as a comma to distinguish between lemmata with the full form of the abbreviated lemma part as either initial or as final component of the term, e.g. *f. caecum* to represent the term *foramen caecum* and *f., jugulêre* to represent the term *jugulêre foramen*.

In this nest the lemma *f. caecum* is a non-grouped sublemma reduced to a partial lemma by means of a process of textual condensation through abbreviation, and alphabetically ordered according to its first component. The lemma *f., jugulère* is a non-grouped sublemma reduced to a partial lemma by means of a process of textual condensation through abbreviation, and alphabetically ordered according to its non-initial component. These lemmata are part of a non-grouped nest with sublemmata reduced to partial lemmata by means of a process of textual condensation through abbreviation with a hybrid alphabetical ordering that allows ordering according to either the first or the last component of the lemma and with examples of both types prevailing in the nest.

### 5.5. Different Levels of Sublemmatisation

Within any lexicographic process a dictionary conceptualisation plan needs to be formulated to make provision for a description of the functions, contents and structures of the envisaged dictionary. A vital component of such a dictionary conceptualisation plan is a description of the macrostructure and the lemmatisation procedures to be employed in the dictionary. When opting for macrostructural procedures involving nesting and niching and the subsequent lemmatisation applications lexicographers realise that the specific dictionary will have at least two types of lemmata, i.e. those in default vertical alphabetical ordering and those included as either nested or niched lemmata. These lemmata can be presented as grouped or non-grouped guiding elements of their articles and can be either main or sublemmata and, in the case of sublemmata, either full or partial lemmata.

One of the advantages of a contemplative approach to lexicographic research is that the metalexigrapher can detect procedures in the ongoing lexicographic practice that need to be described so that they can be accounted for within a general theory of lexicography and be made available within a theoretical model for further applications. Therefore a contemplative approach can contribute to the expansion of a general theory of lexicography.

In this paper reference has been made to some innovative macrostructural procedures in the WGW. One of the most salient procedures, a procedure that has not been used in too many dictionaries, if at all, and that has not been described from a theoretical perspective too often, if at all, is the macrostructural procedure of double-layered sublemmata.

By employing a macrostructural procedure of grouped or non-grouped article clusters the lexicographer opts for a double-layered lemmatisation process, with the main lemmata ordered within the default vertical alphabetical arrangement representing the first layer and the grouped or non-grouped lemmata, either main or sublemmata, within the article clusters, representing the second layer. The

WGW takes this procedure of layered lemmatisation a step further by also employing it within the domain of article clusters, i.e. within the second layer of a traditional double layered lemmatisation application.

Attached to the article of the lemma sign *hemostase* there is a nest containing two non-grouped articles headed by the sublemmata *h., eksogene* and *h., endogene*. These two partial lemmata have been ordered according to the non-initial lemma part *hemostase*, abbreviated as *h*. This procedure of adding nested lemmata is an application of the default double-layered lemmatisation procedure of this dictionary. However, attached to the article of the lemma sign *h., endogene* a non-grouped cluster of articles is presented, containing the lemma signs *e.h. deur ekstravaskulêre meganismes*; *e.h. deur intravaskulêre meganismes* and *e.h. deur vaskulêre meganismes*. The terms represented by these sublemmata have the term presented by the preceding higher level sublemma *h., endogene* as initial component. The multiword terms *endogene hemostase deur ekstravaskulêre meganismes*, *endogene hemostase deur intravaskulêre meganismes* and *endogene hemostase deur vaskulêre meganismes* are included as sublemmata attached to the article of the lemma representing the term *endogene hemostase*, which in its turn is attached to the article of the lemma sign *hemostase*. In this partial article stretch the lemma *h., endogene* is a sublemma of the lemma *hemostase* and the lemmata *e.h. deur ekstravaskulêre meganismes*; *e.h. deur intravaskulêre meganismes* and *e.h. deur vaskulêre meganismes* are sublemmata of the sublemma *h., endogene*. This is the result of an innovative procedure of double-layered sublemmata. In this application the sublemma of the first layer is ordered according to a non-initial lemma part with each sublemma of the second layer ordered according to their first lemma part, which is the non-initial part of the first layer sublemma.

Paragraph 5.2 offers a discussion of the occurrence of macrostructurally-isolated sublemmata, i.e. when there are only two terms in a group and the lemmatisation leads to one of them being presented as main lemma and the other as a sublemma. The WGW also employs the procedures of macrostructurally-isolated sublemmata and double-layered sublemmata in a single partial article stretch. This is seen in the partial article stretch attached to the article of the lemma *tomografie*. A single macrostructurally-isolated sublemma *t., rekenaar-*, representing the multiword term *rekenaartomografie*, and its article is attached to the article of the lemma *tomografie* as an application of the default double-layered lemmatisation procedure, but with only one lemma and its article in this second layer of lemmata. The sublemma *t., rekenaar-* functions as entrance lemma to the articles of a nest containing no less than nine further sublemmata, including *r., aksiale*, *r., dinamiese* and *r., ultrasono-*, representing the multiword terms *aksiale rekenaartomografie*, *dinamiese rekenaartomografie* and *ultrasonorekenaartomografie*. These terms are lemmatised as partial lemmata, functioning in a procedure of double-layered sublemmata as

sublemmata of the first layer sublemma *t.*, *rekenaar*-, which in its turn also is a macrostructurally-isolated sublemma.

Another application of the procedure of double-layered sublemmata is seen in the partial article stretch attached to the article of the lemma sign *sifilis*. In a non-grouped partial article stretch of ten nested articles, ordered in terms of the non-initial part of their partial lemmata, the last two partial lemmata are *s.*, *serologiese toetse vir* and *s.*, *tersiêre*. As happens quite often in the WGW where an abbreviation of a term is given, the lemma sign *s.*, *serologiese toetse vir* is immediately followed by an item giving its abbreviation in brackets, i.e. (*STS*). The sublemma *s.*, *serologiese toetse* becomes the entrance lemma to a second layer of sublemmata that have the abbreviation *STS*, introduced in the article of the first layer sublemma, as mutual component and as lemma part that determines the macrostructural position of these sublemmata: *STS*, *sifting*- and *STS*, *spesifieke*.

By opting for a procedure of double-layered sublemmata the lexicographer introduces a transitive relation between the different layers of sublemmata. This adds to the extent of bringing terms with mutual components together and linking them explicitly.

One problematic issue from the perspective of the dictionary user is the fact that it is not quite clear whether the abbreviation used in these two sublemmata must be interpreted as being the form occurring in the actual term or only the form employed in the condensed item giving the lemma sign.

## 6. CONCLUSION

Using the general theory of lexicography to plan a dictionary gives lexicographers the opportunity to opt for existing lexicographic procedures or new procedures, but also to combine existing and new procedures by using existing procedures as a basis for innovative adaptations and applications. In terms of macrostructural procedures the lexicographers of the WGW have done exactly that. The way in which numerous multiword terms have been entered into the macrostructure by means of a variety of procedures resulted in different types of sublemmata. The fact that multiword terms often prevail in semantically related groups has increased the need for the use of sublemmata. Term formation uses existing terms as a basis for new terms and these complex new terms then form a basis for a second expansion. These different layers in the structure of terms also create the opportunity for different layers of sublemmata. In this regard the WGW has applied existing theoretical guidelines in an innovative way to ensure the best possible and wide-ranging macrostructural coverage. Here a contemplative approach to lexicography offers the basis for innovative transformative approaches.



## REFERENCES

### Dictionaries

- Brink, A.J. & J. de V. Locher (Eds.) 2011. *Woordeboek vir die Gesondheidswetenskappe/Dictionary for the Health Sciences*. Cape Town: Pharos.
- Gouws, R. et al. (Eds.) 2010. *Grondslagfasewoordeboek*. Cape Town: Maskew Miller Longman.
- Martin, W. et al. (Eds.) 2011. *Groot woordenboek Afrikaans en Nederlands*. Houten: Prisma.
- Wiesner, E. & R. Riibeck. 1991. *Wörterbuch der Veterinärmedizin*. Stuttgart.

### Other Literature

- Bergenholtz, H. & R.H. Gouws. 2007. The Access Process for Fixed Expressions. *Lexicographica*, 23:236-260.
- Bergenholtz, H., S. Tarp & H.E. Wiegand. 1999. Datendistributionsstrukturen, Makro- und Mikrostrukturen in neueren Fachwörterbüchern. *Fachsprachen. Languages for Special Purposes. An International Handbook of Special-Language and Terminology Research*, edited by L. Hoffmann et al. Berlin, De Gruyter:1762-1832.
- Botha, W.F. 1991. Die lemmatisering van uitdrukkings in verklarende Afrikaanse woordeboeke. *Lexikos* 1:20-36.
- Gouws, L. 2006. *Die bewerking van idiome in tweetalige woordeboeke: 'n hulp vir vertalers?* MPhil thesis. Stellenbosch: University of Stellenbosch.
- Gouws, R.H. 1989. *Leksikografie*. Cape Town: Academica.
- Gouws, R.H. 1996. Idioms and Collocations in Bilingual Dictionaries and their Afrikaans Translation Equivalents. *Lexicographica* 12:54-88.
- Gouws, R.H. 2010. Fixed word combinations as second level treatment units in dictionaries. *Feste Wortverbindungen in Wörterbücher*, edited by Durco, P. Hildesheim: Georg Olms Verlag.51-63.
- Gouws, R.H. to be published. Establishing and developing a dictionary culture for specialised lexicography.
- Gouws, R.H. & D.J. Prinsloo 2005. *Principles and Practice of South African Lexicography*. Stellenbosch: SUN Press.
- Prinsloo, D.J. 1994. Lemmatization of verbs in Northern Sotho. *South Africa Journal of African Languages* 14(2):93-102.

- Prinsloo, D.J. & R.H. Gouws. 1996. Formulating a new dictionary convention for the lemmatization of verbs in Northern Sotho. *South African Journal of African Languages* 16(3):100-107.
- Tarp, S. 2000. Theoretical challenges to LSP lexicography. *Lexikos* 10:189-208.
- Wiegand, H. E. 1989. Aspekte der Makrostruktur im allgemeinen einsprachigen Wörterbuch: alphabetische Anordnungsformen und ihre Probleme. *Wörterbücher. Dictionaries. Dictionnaires. An International Encyclopedia of Lexicography*, edited by Hausmann, F.J., Zgusta, L., Reichmann, O. & H.E. Wiegand. Berlin: De Gruyter.371-409.
- Wiegand, H. E. 1998. Altes und Neues zur Makrostruktur alphabetischer Printwörterbücher. *Wörterbücher in der Diskussion III*, edited by H.E. Wiegand. Tübingen: Max Niemeyer.348-372.
- Wiegand, H. E. 2010. Zur Methodologie der Systematischen Wörterbuchforschung: Ausgewählte Untersuchungs- und Darstellungsmethoden für die Wörterbuchform. *Lexicographica* 26:249-330.
- Wiegand, H.E. & R.H. Gouws. 2011. Theoriebedingte Wörterbuchformprobleme und wörterbuchformbedingte Benutzerprobleme I. *Lexikos* 21:232-297.
- Wiegand, H.E. & R.H. Gouws. 2013. Macrostructures in printed dictionaries: An overview. Dictionaries. Recent developments with special focus on computational lexicography. *Supplementary volume to Wörterbücher. Dictionaries. Dictionnaires. An International Encyclopedia of Lexicography*, edited by R.H. Gouws, U. Heid, W.Schweickard & H.E. Wiegand. Berlin: De Gruyter.

---

## **PART 4:**

### HLT RESEARCH AND DEVELOPMENT

---

## CHAPTER 16

# FREQUENCY-BASED DATA SELECTION FOR STATISTICAL MACHINE TRANSLATION WITH SCARCE RESOURCES

Cindy A. McKellar and Hendrik J. Groenewald

*Centre for Text Technology, North-West University, Potchefstroom, South Africa*  
Cindy.McKellar@nwu.ac.za

### 1. INTRODUCTION

Machine translation is the process where text is automatically translated from one language to another by computer software. Interest in machine translation as an area of research started to gain momentum with the advent of the Cold War when the USA and the Soviet Union realised that they both had the need for fast and accurate translation of documents between Russian and English (Craciunescu *et al.* 2004). The complexity of machine translation was however grossly underestimated in those early years. This, together with the publication of the sceptical ALPAC report (Pierce *et al.* 1966) caused that interest in (and funding of) machine translation research gradually decreased (Hutchins 1996:9). Because of this machine translation was considered to be an impossible task and somewhat of a joke for many years. Urban legends like the machine translated version of Matthew 26:41 "*The spirit is willing, but the flesh is weak*" as "*The whiskey is strong, but the meat is rotten*" became well known (Hutchins 1995:17).

This viewpoint is luckily changing because of the recent advances that are being made in the field of statistical machine translation. These advances can be ascribed to a renewed interest in statistical machine translation as an area of research, which was instigated by increases in computer processing power and the expansion of the internet (and the resulting increase in the availability of parallel corpora) in recent years. Although the output quality of statistical machine translation systems is constantly increasing, there is still much room for improvement. Machine translation therefore remains an unsolved problem and is still considered to be one of the biggest challenges in natural language processing.

The advances that are being made are changing the general perception of machine translation from something that is considered to be a gimmick to a valuable tool for providing access to information in a multilingual environment. For example, Google's free online machine translation service, Google Translate ([translate.google.com](http://translate.google.com)) offers instant automatic translation between more than fifty languages in any direction. The main advantage of Google Translate is that it

assists the user to understand the general meaning of texts that are only available in languages that are foreign to the user.

The purpose of machine translation is not only to provide access to information that is exclusively available in a foreign language, the translation industry is also becoming increasingly aware of the positive contribution that machine translation could have in assisting the human translator. This is evident from the fact that well known computer-assisted software suites such as *SDL Trados* ([www.trados.com](http://www.trados.com)) and *OmegaT* ([www.omegat.org](http://www.omegat.org)) currently offer both translation memories and machine translation as translation aids. Here the advantage of machine translation is that it can provide a pre-translation of a text segment in cases where a 100% match or high fuzzy match (70%-99%) is not available in the translation memory.

In South Africa with its eleven official languages, there is a large demand for translation services and access to information in all eleven languages. Machine translation can consequently play an important role in the South African context by promoting access to information and assisting the human translator. The South African Government is aware of the positive impact that machine translation may have in this regard and is therefore involved in a number of initiatives and projects to develop various machine-aided translation tools and resources (including machine translation systems) for South African language pairs.

One such a project is the Autshumato Project ([www.autshumato.sourceforge.org](http://www.autshumato.sourceforge.org)) that was commissioned in 2006 by the South African Department of Arts and Culture (Groenewald & Du Plooy 2010:27). The aim of the Autshumato Project was to develop open source machine-aided translation tools and resources for South African languages, including machine translation systems for three South African language pairs, namely English-Afrikaans, English-isiZulu and English-Sepedi.

Despite the lack of required resources such as parallel corpora, it was decided that statistical machine translation would form the basis for the development of the three machine translation systems in the Autshumato Project for the following reasons (Groenewald & Du Plooy 2010:28):

- Statistical machine translation is currently the preferred approach of numerous industrial and academic research groups.
- State-of-the art open source statistical machine translation toolkits are readily available.
- Less expert linguistic knowledge is required to create a working baseline system in comparison to rule-based machine translation systems.

Statistical machine translation requires at least two resources, namely an aligned, bilingual corpus and a monolingual corpus in the target language (the language into which the text is to be translated). The quality of the machine translation

output depends on both the quality and the quantity of the available data. According to Och (2005), a solid base for developing a statistical machine translation system from scratch would consist of having a parallel text corpus of at least one million words and a monolingual corpus in the target language consisting of more than a billion words. In the case of resource-scarce languages, like the indigenous South African languages, this requirement can be problematic as both parallel and monolingual corpora are not readily available.

The lack of parallel corpora implies that alternative methods had to be sought in order to improve the quality of the machine translation systems that were developed in the Autshumato Project. This includes augmenting the statistical machine translation output with rules based on expert linguistic knowledge, manipulation of the training data and the application of data selection techniques. The data selection techniques aimed to select the training data that would result in a better learning rate than the random selection of training data. One method for the selection of training data is to base the data selection process on word and phrase frequency, which forms the central theme of this chapter.

The rest of this chapter is organised as follows: a general introduction to statistical machine translation is provided in the next session. This is followed by an overview of related work in Section 3. Section 4 focuses on frequency-based data selection, while Section 5 gives information about the experimental setup. The results of the various machine translation systems are presented in Section 6 and an interpretation of the results is provided in Section 7. The chapter concludes in Section 8 with some directions for future work.

## 2. STATISTICAL MACHINE TRANSLATION

Any translation must have the following two qualities:

- The meaning of the translated sentence must be faithful to the meaning of the original sentence.
- The translated sentence must be a natural sentence in the target language.

The purpose of machine translation can therefore be viewed as the maximisation of a function that encompasses both these qualities (Jurafsky & Martin 2009:911). If a translation is performed from a source language ( $S$ ) to a target language ( $T$ ), this function can be presented as

$$\tilde{T} = \operatorname{argmax} P(S|T)P(T) \quad [1]$$

where  $\tilde{T}$  represents the best possible translation of the source sentence  $S$ . Equation 1 shows that two components are required for statistical machine translation, namely a translation model  $P(S|T)$  that represents the probability that the target

string is the translation of the source string and a language model  $P(T)$ . The language model represents the probability that the translated sentence occurs as a natural sentence in the target language. The translation model is computed on the basis of a parallel corpus, while the language model is generated with the aid of a monolingual target language corpus.

The statistical machine translation systems described in this contribution were developed with the open source Moses Statistical Machine Translation Toolkit (Koehn *et al.* 2007). Moses facilitates the development of statistical machine translation systems for any language pair. Only two basic resources are required, i.e. a parallel corpus for the training of the translation model and a monolingual corpus in the target language for the training of the language model. The language models employed in this research were trained with the SRI Language Modelling Toolkit (Stolcke 2002:901).

Moses offers two types of translation models, i.e. phrase-based and tree-based models. The phrase-based translation model is currently the dominant approach in statistical machine translation (Goutte 2009:18) and is also the approach employed in the experiments presented in this chapter. Phrase-based machine translation operates at the phrase level and consists of three basic steps (Jurafsky & Martin 2009: 913):

- Grouping the source words into phrases.
- Translating each source phrase into a target phrase according to a phrase-based translation model.
- Reordering of the target phrases.

The reordering of the phrases in the target language is performed according to a distortion model, which provides a measure of the distance between the positions of the source language phrase and the target language phrase. Both the translation and the distortion models are computed on the basis of a parallel corpus.

### **3. RELATED WORK**

As mentioned in Section 1, the output quality of a statistical machine translation system depends to a large extent on the size of the parallel corpus that is used as training data. Merely adding more training data is therefore a common strategy for improving the output quality of a statistical machine translation system. This is explained by the fact that the more training data is available, the more opportunities the training algorithm has to observe how certain words and phrases are translated.

Parallel training data can unfortunately be difficult to obtain, especially when working with resource-scarce languages. One method to overcome a lack of parallel data is to generate parallel data through the selection of monolingual sentences (in one of the two languages for which the machine translation system is being trained for), which can be translated by professional translators. Since manual translation is such a time consuming and expensive process, it is important to select “informative” monolingual sentences that will result in a higher learning rate than randomly selected data. An informative sentence is a sentence that contributes new knowledge to the machine translation system. This knowledge can, for example, include words or phrases that are not in the translation model of the system yet. This process of manipulating the selection process to select informative training data is a common practise in the field of machine learning and is known as active learning. Active learning is the direct opposite of passive learning, which is the random selection of training data. Active learning methods usually achieve better results than passive learning methods (Tong 2001:5). Several studies (Haffari *et al.* 2009:415; Kato & Barnard 2007:1; Mandal *et al.* 2008:261) have shown that data selection methods based on active learning techniques can also lead to better results when applied to data selection for machine translation.

Haffari *et al.* (2009:415) used an active learning method to improve the output of a phrase-based statistical machine translation system. They started with a small parallel corpus and a larger monolingual corpus in the source language. The parallel corpus was used to train a machine translation system, which was then used to translate the monolingual corpus. The translated version of the monolingual corpus was then used together with the parallel corpus to train a new machine translation system, which was used to translate a separate test data set for evaluation purposes. In the next step the most informative sentences were removed from the monolingual corpus and manually translated. The translated sentences were then appended to the original, small parallel corpus (Haffari *et al.* 2009:416) and the entire process was iteratively repeated until a desired level of translation quality (measured by the BLEU score) was reached.

Various methods for the selection of informative sentences for manual translation were evaluated and compared to a baseline system (trained with randomly selected training data) in order to determine the method that yielded the best results (Haffari *et al.* 2009:416). The first data selection method was based on the frequencies of phrases and  $n$ -grams in both the parallel and monolingual data. This selection method is based on the assumption that frequently occurring phrases in the monolingual data needs to be included in the parallel training data. Phrases that already occur frequently in the parallel training data are less important because their translations are already available (Haffari *et al.* 2009:417). Other data selection methods that were investigated included an adapted version of the Hierarchical Adaptive Sampling-algorithm and a method that calculated the



amount of information that got lost if the direction of translation was reversed (Haffari *et al.* 2009:418). It was found that most of the investigated data selection methods gave better results than random data selection (Haffari *et al.* 2009:420).

Kato and Barnard (2007:1) used a combination of active and semi-supervised learning techniques to select additional training data for statistical machine translation. Their approach was developed with the assumption that a small parallel corpus and a larger monolingual corpus (in the source language) are available. First a preliminary machine translation system was trained with the available parallel corpus and then used to translate the monolingual corpus to the target language (Kato & Barnard 2007:2). The translated data was then separated into two parts according to a certain threshold: one part containing the data with high confidence scores (a measure of the “confidence” or “certainty” of each translation) and the other part, the data with low confidence scores. The data with the low confidence scores was manually translated and then added to the bilingual data. The number of instances that were manually translated was determined according to the separation threshold, which is in turn determined by the capacity of the human translator or by the performance of the current model. The new training data was used to train a new machine translation system which was used to translate the remaining monolingual data (the data with the high confidence scores) (Kato & Barnard 2007:2). The entire process was iteratively repeated as long as the results kept improving (Kato & Barnard 2007:2). The experiment was carried out on four South-African language pairs namely English-Afrikaans, English-isiZulu, English-Setswana and English-isiXhosa (Kato & Barnard 2007:2). The application of active and semi-supervised learning on data selection did cause an improvement over the random baseline for all the language pairs (Kato & Barnard 2007:4). The largest improvement occurred with the language pair English-Setswana where the BLEU score improved with 0.12 (Kato & Barnard 2007:3).

Another study (Mandal *et al.* 2008:261) experimented with two different data selection methods, as well as a combination of the two methods to determine the effect on machine translation for the language pair Mandarin-English. The first method attempted to choose the most informative sentences, based on the differences between translations made by different machine translation systems (Mandal *et al.* 2008:261). Three machine translation systems were trained on the same data and used to translate the monolingual source language text. An automatic evaluation metric was used to compare the three translations (each time using the other two as references) after which a linear regression function was trained to predict the translation quality (TER evaluation metric) on new, unseen, source language text (Mandal *et al.* 2008:262). The sentences with the highest predicted TER scores were chosen to be translated and added to the training data (Mandal *et al.* 2008:263).

The second data selection method made use of two language models, one that was trained on only the source language sentences of the bilingual corpus and the other one was trained on all available source language sentences. The purpose of the two language models was to identify data that occurs rarely in the training data, but frequently in the monolingual data (Mandal *et al.* 2008:262). The sentences chosen for manual translation were selected according to a perplexity ratio measure that was low if the sentence occurred rarely in both the monolingual and parallel corpora. A low perplexity ratio meant that the sentence was unlikely to increase the learning rate of the machine translation system and would therefore not be chosen for manual translation (Mandal *et al.* 2008:262). A high perplexity ratio meant that the involved sentence was scarce in the bilingual data, but frequent in the monolingual data and therefore needed to be selected for manual translation and included in the bilingual training data (Mandal *et al.* 2008:262).

The two different data selection methods were also combined by selecting an equal amount of data with each method to create a hybrid system. Mandal *et al.* (2008:262) found that by using these data selection methods it is possible to achieve good results with less data than would be needed with random selection. In some cases these data selection methods even led to machine translation systems that achieved better results than systems trained with all available data.

#### **4. FREQUENCY-BASED DATA SELECTION**

Studies show that a relatively small part of a language's vocabulary accounts for a large part of the general language usage (Nation 2001:13) and that the majority of a language's vocabulary is therefore rarely used. According to Nation (2001:15) as little as 5,000 words can account for up to 89.4% of general English usage. This leads to the deduction that in order to improve the quality of a machine translation system as quickly as possible, sentences containing frequently occurring words must be added to the training corpus of the machine translation system before sentences containing less frequently occurring words are added. This approach ensures that the machine translation system first learns how to translate the frequently occurring words that account for the majority of language usage, while the translations of scarcer words are learned at a later stage. This approach forms the basis of the research presented here, since the main aim of this chapter is to show that in cases where small amounts of parallel corpora (<50,000 source/target words) is available, this approach of first adding data that contains frequently occurring words to the training data will lead to higher learning rates in comparison to the situation where additional training data is randomly selected.

The selection of training data that contains frequently occurring words depends on the availability of lists of frequently occurring English *n*-grams. For the

experiments presented in this chapter the lists were automatically extracted from the English half of the German-English *Europarl* corpus (Koehn 2005:79). The *Europarl* corpus is a very large corpus that is freely available on the web. All  $n$ -grams occurring ten or less times in the *Europarl* corpus were removed from the lists to eliminate possible spelling or typing errors. Subsequently, five lists (1-gram, 2-gram, 3-gram, 4-gram and 5-gram) were used during the data selection process. Each list was sorted according to frequency from high to low.

Next a method for scoring each sentence according to the frequency of the content was developed. The score was formulated in such a way that frequently occurring  $n$ -grams caused a bigger increase in the score than the less frequently occurring  $n$ -grams. The effect of this was that a sentence that contained frequently occurring  $n$ -grams was assigned a higher score than a sentence that contained less frequently occurring  $n$ -grams. A higher score therefore indicated a sentence consisting of frequently occurring phrases that had to be translated and included in the parallel training data.

The first step in the scoring process was to assign a score to each of the  $n$ -grams in the lists to indicate how frequently they occur. This “ $n$ -gram score” was calculated by subtracting the position of the particular  $n$ -gram in the frequency-sorted list (sorted from high to low) from the number of  $n$ -grams in the list plus one, and dividing the resulting number by the number of  $n$ -grams in the list.

If the position of an  $n$ -gram in the frequency-sorted list is represented by  $p$  and the number of  $n$ -grams in the list by  $n$ , the score assigned to each  $n$ -gram was calculated as follows:

$$N\text{-gram score} = \frac{n - p + 1}{n} \quad [2]$$

The final frequency score for each sentence/phrase gave an indication of how frequent all the 1-5-grams in the sentence/phrase were. This score was calculated by adding the scores of all the individual  $n$ -grams in the sentence/phrase and dividing this sum by the number of  $n$ -grams as shown in Equation 3. The  $n$ -grams that did not occur in the  $n$ -gram lists contributed zero to the total score.

$$\text{Frequency score per sentence} = \frac{\sum n\text{-gram scores}}{\text{Number of } n\text{-grams}} \quad [3]$$

## 5. DATA SELECTION EXPERIMENTS

The data selection method described in the previous section was applied to the selection of training data for the three statistical machine translation systems that were developed in the Autshumato project (i.e. English-Afrikaans, English-isiZulu

and English-Sepedi) to determine if it could lead to increased learning rates. Although the three target languages are considered to be resource-scarce languages, the amounts of parallel corpora that are freely available for the three language pairs are significantly higher than what is freely available for the other official South African languages.

We started all our data selection experiments with machine translation systems trained with training data consisting of 50,000 source words. To evaluate the effect of the data selection methods under investigation, additional training data was selected and appended to the original training data. In order to keep the experiments for the three language pairs comparable, the additional training data was appended in batches of 50,000 words in nine iterations, which resulted in systems trained with parallel training data consisting of 500,000 words in the final iteration. The word count was performed on the English half of the parallel data and includes punctuation.

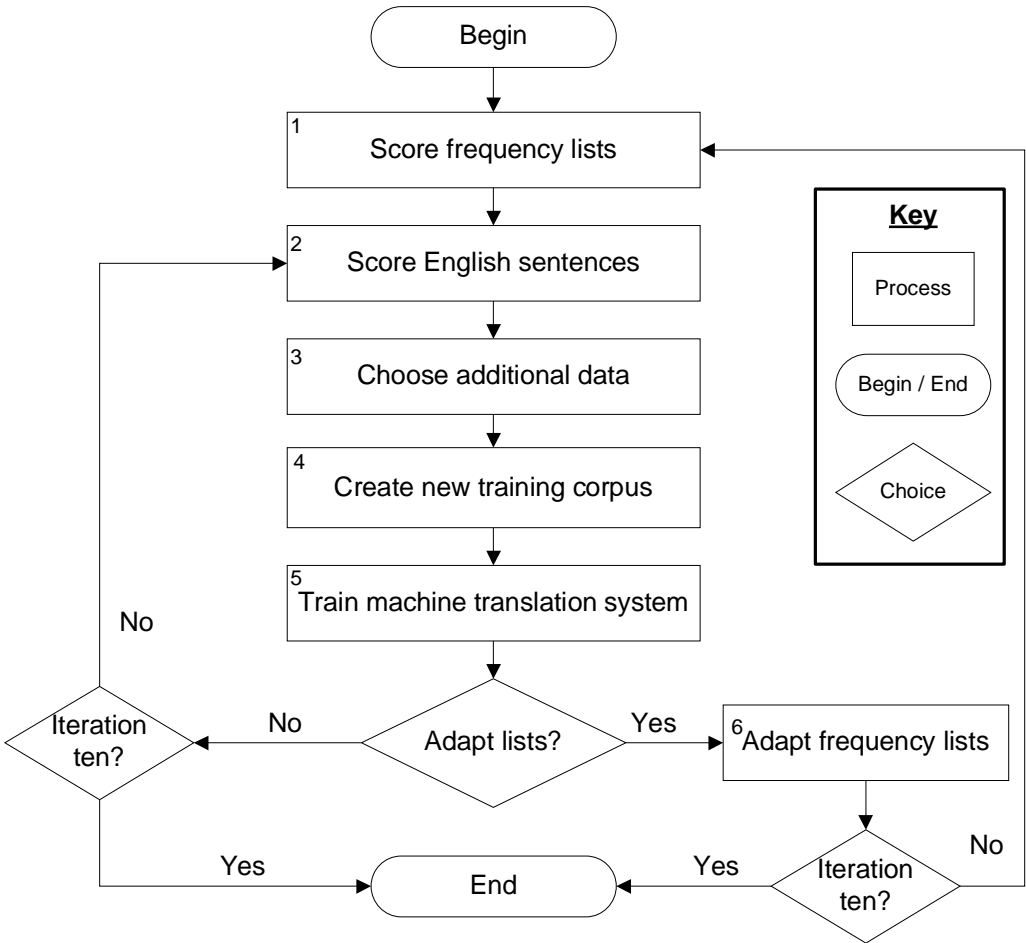
In the real-world scenario of scarce training data, the additional English data will be selected and then manually translated before appending it to the existing bilingual parallel training data. In our experiments the additional data was not manually translated, because we already had parallel corpora consisting of more than 500,000 words available for each of the three language pairs. We therefore merely extracted the translations of the additional selected English text units from the target side of the available parallel corpora before appending it to the training data.

Although it is possible to keep selecting sentences based on the same  $n$ -gram frequency lists in every iteration, this approach will continue to select sentences containing the most frequent  $n$ -grams regardless of whether those  $n$ -grams are already present in the training data or not. This can lead to unnecessary duplication in the training data. For this reason a second data selection experiment was performed during which the frequency lists were adapted after each iteration by removing all the  $n$ -grams already present in the training data from the  $n$ -gram frequency lists. So instead of assigning the most weight to the most frequent  $n$ -grams, whether they were already present in the training data or not, those  $n$ -grams that were already in the training data contributed nothing to the sentence's frequency score. This caused the focus of the data selection process to shift to select the sentences that contained the most frequent  $n$ -grams that were not yet available in the training data.

The flowchart in Figure 1 gives an overview of both data selection methods described in this section. In the first step, scores were assigned to the  $n$ -grams in the frequency lists according to Equation 2. The scored frequency lists were then used to assign a frequency score (see Equation 3) to each English sentence in the corpus of possible additional training data (Step 2). The English sentences with the highest

scores, along with their target language translations, were then removed from the corpus of existing bilingual aligned sentences/phrases (Step 3) and appended to the training data from the previous iteration (if any) in order to create a new training corpus (Step 4). The new training corpus was then used to train a new machine translation system (Step 5). In the case of the method where the  $n$ -gram frequency lists were adapted after each iteration, the  $n$ -gram frequency lists were adapted, by removing all the  $n$ -grams that already occurred in the training data (Step 6). If a further iteration was needed, the frequency lists were then scored anew (Step 1) and the next iteration started. In the case of the data selection method where the frequency lists did not need to be adapted and rescored, the program proceeded directly to Step 2 in the next iteration.

**Figure 1:** Flowchart of data selection process



## 6. EVALUATION

The execution of the various data selection experiments entailed the training of a large number of machine translation systems. All parameter settings were kept constant during the training of the different machine translation systems to ensure that comparable results were obtained. These systems were all evaluated with automatic evaluation metrics to determine the effect of data selection on the quality of the machine translated texts. The BLEU (Papineni *et al.* 2002:311) and NIST (Doddington 2002:138) evaluation metrics were used for the evaluation. Both evaluation metrics are based on the principle that a “good” machine translated text is one that shows a high degree of similarity to a human translation. The similarity between the translated text and the reference translation can be measured according to the number of overlapping  $n$ -grams; the higher the number of overlapping  $n$ -grams, the better the quality of the machine translated text and the higher the NIST and BLEU scores. Since any source language sentence can have multiple correct translations, even human translations seldom achieve the highest possible NIST and BLEU scores. It is therefore advisable to use multiple different human translations of the same source text as reference translations. For this evaluation we used a source text consisting of 200 sentences, with four reference translations performed by professional, accredited language practitioners.

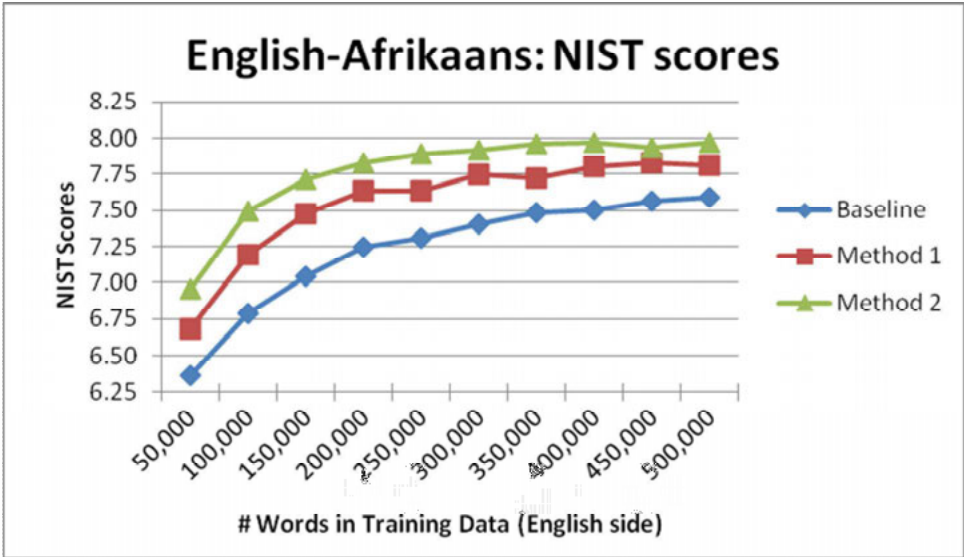
In order to determine if frequency-based data selection caused an improvement in machine translation quality, the results of the machine translation systems had to be compared to a baseline. The baseline was created by randomly selecting parallel training data consisting of 50,000 words in 10 iterations (the same number of words and iterations as selected in the other experiments). The baseline experiment was repeated five times for each language pair to reduce the possibility of sampling errors. The average NIST and BLEU scores of the five runs were used as the baseline scores.

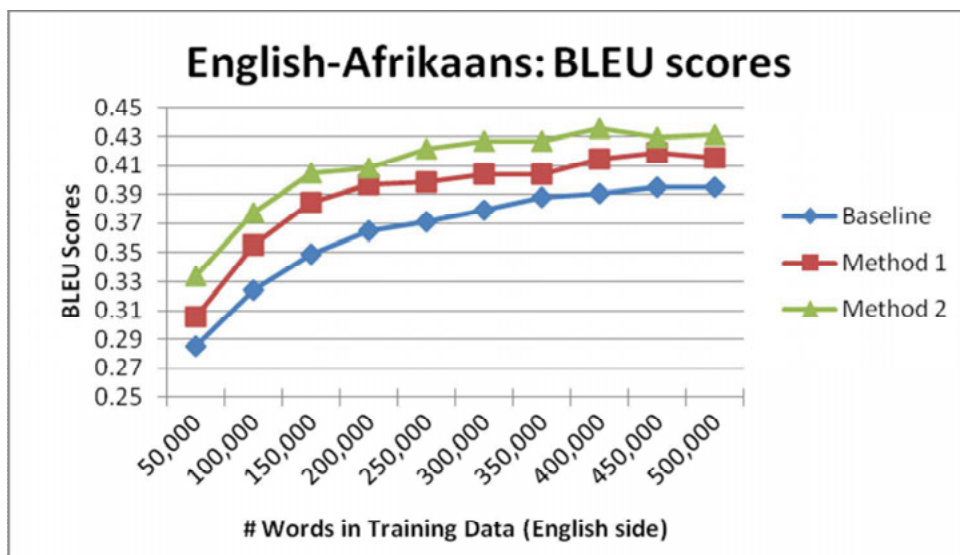
The results of the automatic evaluation are provided in the rest of this section. Tables 1, 2 and 3 show the NIST and BLEU scores obtained by the three machine translation systems. Each table contains the scores for the ten iterations (“ITR”) of the baseline and the two data selection experiments, one without the adaption of frequency lists (Method 1) and the other with the adaption of frequency lists (Method 2). The information displayed in the tables is also graphically represented for comparison purposes in Figures 2-7. Figure 2 shows a graph of the NIST scores of the three machine translation systems for the language pair English-Afrikaans, while Figure 3 displays the BLEU scores for the same three systems. Figures 4 and 5 show the graphs for the English-isiZulu systems and Figures 6 and 7 those of the English-Sepedi systems.

Table 1: NIST and BLEU scores for English-Afrikaans

	ITR 1	ITR 2	ITR 3	ITR 4	ITR 5	ITR 6	ITR 7	ITR 8	ITR 9	ITR 10
NIST										
Baseline	6.3671	6.7878	7.0454	7.2419	7.3074	7.4008	7.4887	7.4981	7.5637	7.5889
Method 1	6.6805	7.1967	7.4786	7.6338	7.6301	7.7475	7.7206	7.7984	7.8311	7.8133
Method 2	6.9522	7.4907	7.7140	7.8249	7.8873	7.9160	7.9573	7.9742	7.9365	7.9701
BLEU										
Baseline	0.2856	0.3240	0.3491	0.3654	0.3711	0.3791	0.3878	0.3903	0.3949	0.3954
Method 1	0.3054	0.3549	0.3846	0.3970	0.3982	0.4041	0.4037	0.4146	0.4194	0.4154
Method 2	0.3342	0.3770	0.4051	0.4080	0.4213	0.4270	0.4270	0.4361	0.4295	0.4317

Figure 2: NIST scores for English-Afrikaans



**Figure 3:** BLEU scores for English-Afrikaans**Table 2:** NIST and BLEU scores for English-isiZulu

	ITR 1	ITR 2	ITR 3	ITR 4	ITR 5	ITR 6	ITR 7	ITR 8	ITR 9	ITR 10
<b>NIST</b>										
<b>Baseline</b>	2.6062	3.0397	3.2510	3.4865	3.6278	3.8259	3.9422	3.9747	4.0824	4.1119
<b>Method 1</b>	3.1178	3.5554	3.7956	4.0733	4.2728	4.1304	4.4239	4.3933	4.6386	4.4621
<b>Method 2</b>	3.3124	3.9343	4.0918	4.3354	4.3081	4.5770	4.5995	4.6294	4.6406	4.6686
<b>BLEU</b>										
<b>Baseline</b>	0.0917	0.1145	0.1211	0.1385	0.1456	0.1586	0.1635	0.1658	0.1722	0.1756
<b>Method 1</b>	0.1140	0.1420	0.1580	0.1740	0.1848	0.1813	0.2007	0.2012	0.2122	0.2099
<b>Method 2</b>	0.1239	0.1588	0.1632	0.1797	0.1817	0.1934	0.1980	0.2116	0.2114	0.2058



Figure 4: NIST scores for English-isiZulu

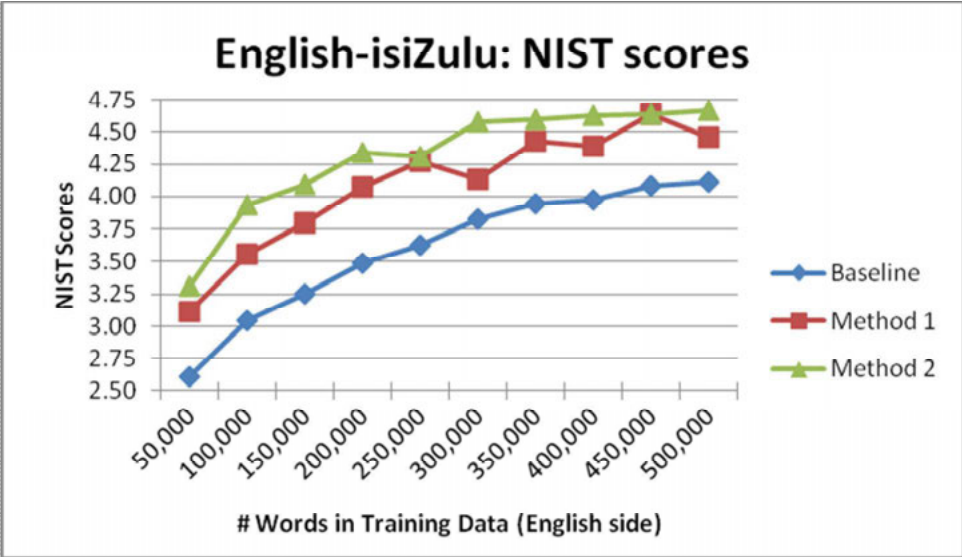
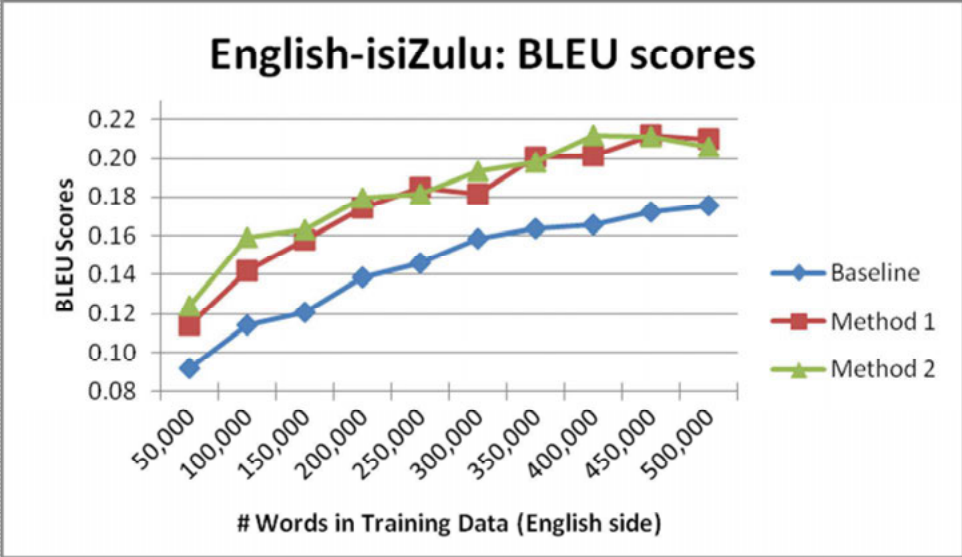
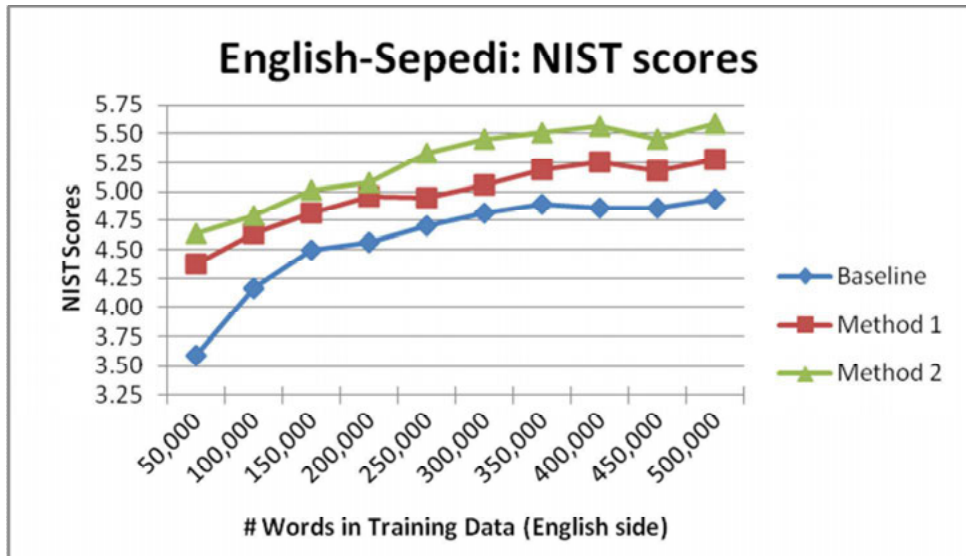


Figure 5: BLEU scores for English-isiZulu

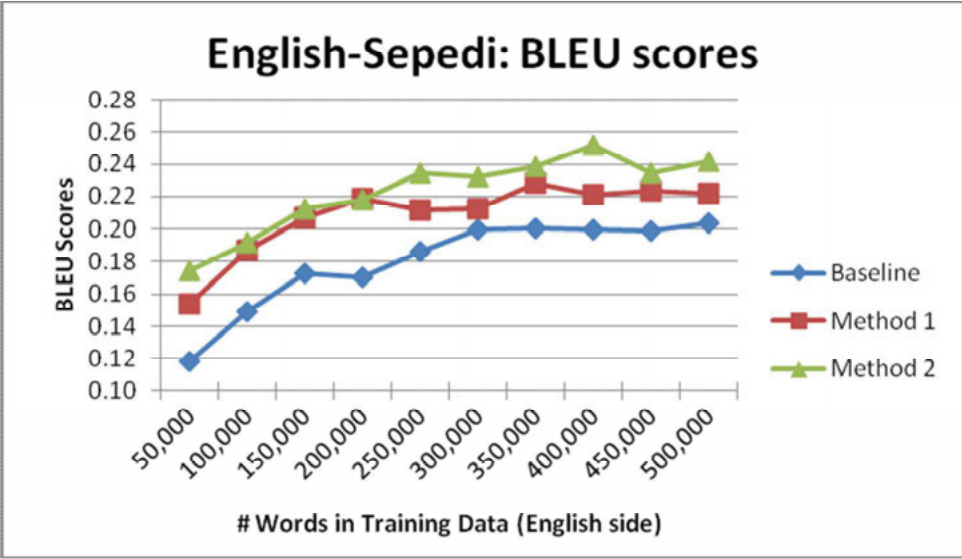


**Table 3:** NIST and BLEU scores for English-Sepedi

	ITR 1	ITR 2	ITR 3	ITR 4	ITR 5	ITR 6	ITR 7	ITR 8	ITR 9	ITR 10
<b>NIST</b>										
<b>Baseline</b>	3.5873	4.1641	4.4967	4.5621	4.7103	4.8194	4.8967	4.8637	4.8545	4.9389
<b>Method 1</b>	4.3790	4.6388	4.8142	4.9539	4.9520	5.0619	5.1906	5.2529	5.1846	5.2756
<b>Method 2</b>	4.6375	4.7920	5.0093	5.0766	5.3321	5.4577	5.5098	5.5701	5.4520	5.5886
<b>BLEU</b>										
<b>Baseline</b>	0.1180	0.1489	0.1730	0.1704	0.1862	0.1998	0.2001	0.1999	0.1988	0.2035
<b>Method 1</b>	0.1540	0.1874	0.2070	0.2188	0.2117	0.2123	0.2285	0.2210	0.2233	0.2220
<b>Method 2</b>	0.1740	0.1920	0.2124	0.2182	0.2347	0.2320	0.2388	0.2516	0.2345	0.2417

**Figure 6:** NIST scores for English-Sepedi

**Figure 7:** BLEU scores for English-Sepedi



**7. ANALYSIS**

The results presented in the previous section show that both data selection methods (Method 1 and Method 2) obtained results that were significantly better than the baseline for all three language pairs under investigation. Both data selection methods achieved higher BLEU and NIST scores than the baseline in the finale iterations of the data selection processes, while the learning rates of the two systems were also better than those of the baseline.

When taking a closer look at the performance of Method 1 vs. Method 2, it is noticed that Method 2 in general outperformed Method 1 in all experiments, except in the case of iteration 4 of the English-Sepedi system and iterations 5, 7, 9 and 10 of the English-isiZulu system, measured according to the BLEU score. In all of these cases the NIST score for the corresponding iterations was not in agreement with the result indicated by the BLEU scores. The results in Table 2 and 3 also show that the amounts by which Method 1 outperformed Method 2 in every iteration was very small (the maximum difference of 0.0041 BLEU points occurs in iteration 10 of the English-isiZulu system). In such cases where the BLEU and NIST scores are not in agreement, an alternative way to determine the best performing data selection method is to resort to a human evaluation of the translation quality of the two systems. However, we did not consider it necessary to resort to human evaluation because Method 2 was outperformed by small margins in only five iterations (out of a total of 30 iterations).

The reason why Method 2 outperformed Method 1 is because the use of Method 1 led to duplication in the training data, and it caused the exclusion of less frequent sentences, which may have resulted in high performance gains, from the training data. New “unseen” data contributes new knowledge to the machine translation systems and leads to higher learning rates. Our results are in agreement with the findings of Haffari *et al.* (2009:417) which stated that phrases that already occur frequently in the parallel training data are less important because their translations are already available.

## 8. CONCLUSION

This study attempted to show that the application of data selection methods can lead to a faster gain in machine translation quality. The training data was selected based on the frequency scores of the source sentences (English). Experiments were conducted for three language pairs, English-Afrikaans, English-isiZulu and English-Sepedi.

The results obtained for both data selection methods signify that appending a small initial training data set of 50,000 aligned text units with additional training data containing frequently occurring  $n$ -grams will result in machine translation systems with a learning rate that is significantly higher than that of a baseline system trained with randomly selected data. The practical implication of these results is that it proves that machine translation systems trained with data that was appended with additional training data containing frequently occurring  $n$ -grams can reach a desired level of accuracy in less time and by using fewer resources than systems trained with randomly selected training data. This is a significant and important result in the context of resource-scarce languages where parallel training data needs to be generated through manual translation.

This data selection process can also be applied in situations where machine translation systems need to be developed for languages where large amounts of parallel training data (more than one million aligned text units) is available. In such cases the emphasis will not be on the selection of monolingual sentences that can be utilised for generating additional training data, but rather on creating machine translation systems that obtain the best possible results with the least amount of training data. Using the minimum amount of training data to achieve better or equal results to that of a system trained with all available training data is beneficial because it results in machine translation systems that translate faster and requires less computational resources.

Two of the three machine translation systems investigated in this work translate from English to languages that are of the Sotho and Nguni language groups. Five of the seven remaining resource-scarce official South African languages (e.g.

isiNdebele, isiXhosa, Sesotho, Setswana and siSwati) are part of either the Nguni or Sotho language groups. We therefore believe that the results obtained for isiZulu and Sepedi provide an indication of the results that can be expected for five of the remaining resource-scarce official South African languages for which machine translation systems have not yet been built.

The encouraging results obtained in this study suggest that further research about data selection techniques for creating machine translation systems can be beneficial, especially in the cases of resource scarce languages. One possibility for future work is the utilisation of morphological and syntactical information in the data selection process. This information can be used to ensure that every possible meaning of a word is reflected in the training data and that there are ample examples of all common sentence structures. During this study all the data selection choices were based on the source language, however it is also possible to base the data selection process on the target language. Further experimentation can also be done regarding the *n*-gram lists; it would, for instance, perhaps be beneficial to extract the *n*-gram frequencies from domain specific data. Data selection methods could also be developed to select “bad” or “unsuitable” data to pinpoint data that should be removed from the corpus. A final interesting possibility for future work is to determine whether the data selection methods presented in this study can also be applied to other resource-scarce languages outside the Germanic, Sotho and Nguni language groups.

## ACKNOWLEDGEMENTS

The authors would like to thank Prof. Justus C. Roux for his valuable inputs and support in this research which was conducted as part of the Autshumato Project. The authors also wish to thank the South African Department of Arts and Culture for their financial support of the Autshumato project.

## REFERENCES

- Craciunescu, O., C. Gerding-Salas. & S. Stringer-O’Keeffe. 2004. Machine translation and computer-assisted translation: a new way of translating? *Translation Journal* 8(3).  
<http://translationjournal.net/journal/29computers.htm>. Accessed: 22-03-2011.
- Doddington, G. 2002. Automatic evaluation of machine translation quality using *n*-gram co-occurrence statistics. *Proceedings of the 2nd International Conference on Human Language Technology Research*, edited by M. Marcus. San Diego, California: Association for Computational Linguistics.138-145.

- Goutte, C., N. Cancedda, M. Dymetman. & G. Foster. 2009. *Learning Machine Translation*. Cambridge: The MIT Press.
- Groenewald, H.J. & L. du Plooy. 2010. Processing parallel text corpora for three South African language pairs in the Autshumato Project. *Proceedings of the Second Workshop on African Language Technology*, edited by G. de Pauw, H.J. Groenewald & G.-M. de Schryver. Malta: European Language Resources Association.27-30.
- Haffari, G., M. Roy & A. Sarkar. 2009. Active learning for statistical phrase-based machine translation. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, edited by C. Leacock & R. Wicentowski. Boulder, Colorado: Association for Computational Linguistics.415-423.
- Hutchins, J. 1995. "The whisky was invisible" or Persistent myths of MT. *MT News International* 11:17-18.
- Hutchins, J. 1996. ALPAC: The (In)famous Report. *MT News International* 14:9-12.
- Jurafsky, D. & J.H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Prentice Hall.
- Kato, R.S.M. & E. Barnard. 2007. Statistical translation with scarce resources: a South African case study. *SAIEE Africa research journal* 98(4):26-30.
- Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. *Proceedings of the Tenth Machine Translation Summit*, edited by the Asia-Pacific Association for Machine Translation. Phuket: Asia-Pacific Association for Machine Translation.79-86.
- Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin & E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.
- Mandal, A., D. Vergyri, W. Wang, J. Zheng, A. Stolcke, G. Tur, D. Hakkani-Tur & N.F. Ayan. 2008. Efficient data selection for machine translation. *Proceedings of the 2008 IEEE workshop on Spoken Language Technology*, edited by the Institute of Electrical and Electronics Engineers. Goa, India: IEEE.261-264.
- Nation, I.S.P. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

- Och, F.J. 2005. Statistical Machine Translation: Foundations and Recent Advances. *Tutorial at MT Summit 2005*. Phuket, Thailand. <http://www.mt-archive.info/MTS-2005-Och.pdf>. Accessed: 22-03-2011.
- Papineni, K., S. Roukos, T. Ward & W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the ACL*, edited by the Association of Computational Linguistics. Stroudsburg, PA, USA: ACL.311-318.
- Pierce, J.R., J.B. Carroll, E.P. Hamp, D.G. Hays, C.F. Hockett & A.G. Oettinger. 1966. *Language and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee (ALPAC), National Academy of Sciences, National Research Council, Washington, DC, Publication 1416.
- Stolcke, A. 2002. SRILM - an extensible language modelling toolkit. *Proceedings of the 7th International Conference on Spoken Language Processing*, edited by J.H.L. Hansen & B. Pellom. Denver, Colorado: ISCA Archive.901-904.
- Tong, S. 2001. *Active learning: theory and applications*. PhD dissertation. Palo Alto, Calif.: Stanford University.

## CHAPTER 17

# VOICE USER INTERFACE DESIGN FOR EMERGING MULTILINGUAL MARKETS

Gerhard van Huyssteen<sup>1</sup>, Aditi Sharma Grover<sup>2</sup>, Karen Calteaux<sup>2</sup>

<sup>1</sup>*Centre for Text Technology, North-West University, Potchefstroom, South Africa*

*gerhard.vanhuyssteen@nwu.ac.za*

<sup>2</sup>*Meraka HLT Group, Council for Scientific and Industrial Research (CSIR)*

*asharma1@csir.co.za, Kcalteaux@csir.co.za*

### 1. INTRODUCTION

Socio-economic development in South Africa poses a variety of challenges, such as bridging communication gaps, providing access to information, dealing with pockets of technological and functional illiteracy, and addressing widespread poverty. South Africa is thus a prototypical example of a developing world nation – it remains marginalised in the world economy, and is considered a medium development country in terms of the United Nations Development Programme's (UNDP's) Human Development Index (International Marketing Council of South Africa, 2010). As a result, it offers many opportunities typical of an emerging region, including support for regional infrastructure projects and transferring knowledge and technology, aspects which are noted in the Economic Development in Africa Report 2010 as important avenues for promoting regional development (UNCTD 2010).

The country boasts a progressive Constitution that protects linguistic human rights and provides for the promotion and equitable use of all eleven of its official languages. A National Language Policy Framework adopted by the South African Government in 2003 mandates all government structures, and encourages private enterprises, to provide for the language needs of end-users. A right to access services in a language of choice is explicitly stated in the Framework (DAC 2003). The need for this is underscored by research indicating that there is not a single official language which can be used as a lingua franca across all linguistic communities in the country (PanSALB 2000).

Within this highly multilingual environment, telephones, and cell phones in particular, are a primary means of communication. In 2007 it was estimated that 87.6% of the South African population had access to a cell phone (Gapminder 2007/2008). By 2008 it was estimated that cell phone penetration in South Africa had increased to 92% (Gapminder 2007/2008) and by mid-2009 to 98% (IT News



Africa 2009). In stark contrast, it is estimated that only 9 to 11% of the South African population had access to the Internet in 2009<sup>1</sup>. One could therefore conclude that telephone-based services have the potential to bridge information and communication gaps prevalent in the South African society (Barnard *et al.* 2010) and other similar developing-world contexts (Tucker *et al.* 2004).

Based on the above, our point of departure for this research is that South Africa holds many market opportunities for multilingual (or multilingualised or multi-language) interactive voice response (IVR) systems, with which end-users could get access to information, complete transactions, activate services, make reservations, etc. This has also been the assumption of the African Speech Technology (AST) project ([www.ast.sun.ac.za](http://www.ast.sun.ac.za)), which was at the time a ground-breaking human language technology (HLT) project in South Africa, led by Prof Justus Roux (Roux *et al.*, 2000). In subsequent years, various other research and development projects in South Africa were based on the very same assumption, including the Open Phone project ([www.meraka.org.za/hlt\\_projects\\_ophone.htm](http://www.meraka.org.za/hlt_projects_ophone.htm)) and the Lwazi project ([www.meraka.org.za/lwazi](http://www.meraka.org.za/lwazi)), both led by the Meraka Institute of the Council for Scientific and Industrial Research (CSIR). All of these projects focussed by and large on the development of resources and technologies for automatic speech recognition (ASR) and text-to-speech (TTS) systems, as well as the development of prototypes of telephone-based services for all eleven official languages of South Africa.

As was pointed out in numerous publications from these and other projects, various interesting challenges arise when developing information and communication technologies (ICTs) for multilingual, emerging markets (e.g. Tucker *et al.* 2004; Barnard *et al.* 2008; Nasfors 2007). Designing effective voice user interfaces (VUIs) is even more challenging when very little is known about users' needs and/or preferences when accessing information via technologies such as IVRs (Barnard *et al.* 2008; Sharma Grover & Barnard 2011a). Previous research has indicated the potential of telephone-based services for stimulating socio-economic growth through access to information (Sharma Grover & Barnard 2011b), as well as some of the challenges associated with designing VUIs (Sharma Grover *et al.* 2009; Sharma Grover & Barnard 2011a; Plauché & Nallasamy 2007) for low-literate users, and taking into account dialectal differences and cost, respectively. Yet very few companies in South Africa specialise in multilingual IVR development; the commercial demand for such services seems to be limited; and very little research data is available in this domain. The VUI designer in commercial markets is therefore left in the dark – often having to base design decisions on own intuition.

The aim of our current research is to get a better understanding of business and design issues related to IVRs in a multilingual, emerging market such as South

Africa, in order to shed light on the challenges relating to VUI design for such markets. We were specifically interested in answering the following questions:

1. What are the most important business drivers for implementing multilingual IVRs?
2. If a multilingual IVR is to be implemented, how many languages should be available, and where/how should the language choice be offered?
3. With regard to choice of voice (of the persona), what are current trends in South Africa?
4. With regard to input modality (touch-tone or speech), why is speech not used more often?

Answers to these questions will provide a glimpse into the challenges facing VUI designers in the country, especially in terms of requirements analysis, a design framework and governance. In the next section, we explicate the methodology we followed to answer these questions. In section 3 we discuss the most important business drivers (and hurdles) related to the development of multilingual IVRs in South Africa. Section 4 focuses on some design issues, specifically related to language choice, options for choice of the voice, and input modality. The chapter concludes with a view on other issues and questions to be investigated in future research.

## **2. METHODOLOGY**

Our investigation into the multilingual IVR landscape of South Africa took a triangulated approach that included: (a) basic data collection on numerous South African telephone-based services; (b) interviews with IVR developers; and (c) interviews with companies that typically use IVR-based solutions for customer interaction and services.

We first conducted an impressionistic survey on a number of IVRs in South Africa by calling into these services to gather information relevant to our research questions. We were specifically interested in –

- whether the service has a multilingual offering, and if so, which languages are offered, and where/how in the IVR the language choice is presented;
- the characteristics (male/female; accent) of the voice artist used for the persona; and
- whether the service uses speech (i.e. ASR) or touch-tone (i.e. dual-tone multi-frequency (DTMF) signalling) as input modality.

Since it is not always clearly discernable from a caller's perspective whether IVR or automatic call distribution (ACD) technology is being used, for our purposes we use the term "IVR" as an all-encompassing term to refer to technology that allows

humans to access computer systems (typically for self-service purposes, or to get access to automated information), or call/contact centre agents (or operators, consultants, representatives, etc.) through a telephone by using either speech or keypad input. We limited our survey to IVRs that required some form of input from the caller (e.g. to make a language choice, or to choose from a menu of options the kind of information or service s/he wants), and explicitly excluded services that merely route callers automatically to a contact centre or reception (i.e. “pure” ACDs). For our test battery, we did not make a distinction between call reasons – all systems were treated equally irrespective of whether the caller wants full self-service (e.g. telephone banking, or getting up-to-date flight information), information (e.g. how to apply for a social grant), or wants to speak to an agent (e.g. to change personal details).

We identified six domains for investigating IVRs: transport (airlines, airports, trains, etc.), banking, medical aid funds, telecommunications, government services, and entertainment (e.g. cinemas, satellite television, ticket services, etc.). Our choice of IVRs in each of these domains was meant to be exemplary rather than exhaustive; note that our choice was skewed by the fact that we tried to include at least all speech-based IVRs that we were aware of, and to also include at least a few services that we knew had a multilingual offering. One should also note that it is often quite challenging to gather accurate information for a survey like this, especially in commercial environments. For example, to assess an IVR offering telephone banking, the surveyor has to have an account with (or at least have access to dummy information for) that specific financial service provider before s/he can log on to that service to do an assessment. Since the aim of this survey was to get an impressionistic view of trends in designing IVRs in South Africa, we are not making any statistical claims regarding the IVR landscape in South Africa as a whole – such claims would be better left to exhaustive surveys. Table 1 presents an overview of the six domains and 34 IVRs we investigated.

**Table 1:** *Overview of domains*

Domain	Examples	# of IVRs
Transport	Flight reservations, flight information, general enquiries	7
Banking	Telephone banking, general enquiries	4
Medical aid funds	Customer care, trauma help	5
Telecommunications	Pre-paid services, balance enquiries, general enquiries	6
Entertainment	Ticket reservation, service activation, general enquiries	4
Government services	Information access, reporting of crime or bad service	8
<b>Total</b>		<b>34</b>

The second part of our approach centred on gathering qualitative information through interviews with various role-players in the South African IVR and contact centre industries. VUI design services in South Africa are primarily provided by

companies operating in the contact centre, business process outsourcing (BPO) or ICT infrastructure industries, or companies that offer bespoke software solutions. These companies typically tend to provide clients with solutions that encompass services such as contact centre integration and automation, provision of hosting technology and infrastructure, contact centre operations, and automation of customer interactions with self-service applications. Thus, VUI design tends to become a smaller, ad hoc, value-added, service required to fulfil the larger solutions provided by these companies, as opposed to a specialised VUI design service often found elsewhere in the world.

We conducted detailed interviews with senior developers or managers at four such companies; of these one is a primarily state-owned entity that provides contact centre and BPO services to various government departments and related subsidiaries, another two are South African divisions of multi-national BPO companies providing comprehensive large-scale contact centre and IVR services (amongst other solutions), and the last one is a local South African small to medium sized enterprise (SME) that develops contact centre and IVR solutions.

Thirdly, we also conducted interviews with representatives from two companies in the banking and government services verticals; these companies are typical clients of the above-mentioned contact centre/IVR solutions provider companies. Both these representatives are responsible for the voice channel in their respective companies.

Through these interviews we sought to ascertain various issues and trends pertaining to multilingual IVR design and usage in South Africa. In addition to getting impressions relating to the questions above, we also sought information about the business drivers for multilingual IVR implementation.

For purposes of objectivity, we decided to treat all interviewees as anonymous; in the remainder of the study we will only refer to Respondent 1, 2, etc. We also keep the names of the IVRs called anonymous, since this chapter is not meant to promote or criticise any of the companies investigated.

### **3. BUSINESS DRIVERS FOR MULTILINGUAL IVRS**

Our survey of IVRs brought to light that multilingual IVRs are not as common as one would expect in a country where eleven official languages are constitutionally recognised; for example, in our survey of eight government services IVRs, only four offer service or information in all eleven languages. If one looks at commercial IVRs, it is clear that very few support more than English. In summary, of the 34 IVRs investigated, only nine have a multilingual offering. Given this view of the multilingual IVR landscape in South Africa, one could ask why multilingual IVRs are not more widespread, especially given the high premium placed on

multilingualism in the South African Constitution. What are the most important business drivers for (not) implementing multilingual IVRs? Elsewhere (Calteaux, Sharma Grover & Van Huyssteen 2012), we discuss these business drivers in more detail; suffice it to only mention some of these drivers here, including:

- Improved branding
- Cost saving (e.g. through increased automation)
- Revenue generation (by offering new or value-added services)
- Increased customer satisfaction
- Customer retention (i.e. preservation of existing revenue base)
- Customer delight
- Improved access to information or services
- Increased call centre agent morale (thereby reducing agent attrition)
- Increased agent utilisation
- Improved productivity
- Access to business intelligence for strategic advantage (e.g. call reasons, customer profiling, etc.)
- Opportunities for upselling of products
- Compliance with laws and regulations
- Political motivation (i.e. to have a multilingual offering could also serve a political agenda by (implicitly) subscribing to values of inclusiveness, nation-building, etc.)
- Competitive advantage (i.e. "If my competitor has a multilingual IVR, I should probably have one too.")
- Response to pain-points
- Multi-channel consistency.

#### **4. CHALLENGES FOR MULTILINGUAL VUI DESIGN**

For the second part of this chapter, we turn to more design-related issues pertaining to multilingual IVRs, including the language offering, the voice of the IVR, and the input modality. Our assumption in this section is that a company or organisation has already made a decision to implement a multilingual IVR, taking cognisance of all prerequisites such as development, testing and tuning costs, maintenance and governance, and back-end integration (including integration in the call centre). Now some design choices have to be made. Our aim is to investigate what companies in South Africa are doing in this regard, and to then reflect briefly on some of these options. The aim is neither to provide exhaustive discussions (due to scope limitations), nor to provide solutions with empirical evidence (which is left for future usability research). Instead, we have chosen to list issues that one should take cognisance of (as a designer), or that could serve as stimulation for future research projects.

#### 4.1. Language Offering

The language offering in a multilingual IVR is a complex issue, and is discussed elsewhere (Van Huyssteen, Sharma Grover & Calteaux, submitted) in more detail. For our current research we focussed on how many languages are offered in the investigated IVRs, what those languages are, and what strategies are being used to present the language choice. Based on our investigation, we categorised the choices that were made in the IVRs we investigated in the following categories:

- Cost driven: monolingual (English only)
- Historically driven: bilingual (English and Afrikaans)
- Linguistically driven: four languages (English, Afrikaans, isiZulu and Sesotho) – this option chooses for maximal coverage of as many speakers as possible, while still keeping costs relatively at bay; this approach is also partially in support of the National Language Policy Framework (DAC 2003).
- Demographically driven: five or six languages (additionally isiXhosa in one IVR, and isiXhosa and Xitsonga in another) – we assume that these choices were based on caller demographics
- Externally driven: eleven languages – the choice to include all official languages is driven either by political reasons or regulatory requirements.

Regarding strategies for presenting the language choice in the IVR, we found that the majority of the multilingual IVRs we investigated offered the language choice upfront, directly after the welcoming message (that was always presented in English). In two cases – both from the telecommunications domain – the language choice was offered upfront on the first call, and subsequently handled by calling line identification (CLI; i.e. the IVR associated the initial language choice with the telephone number). In one of these IVRs, the opportunity to change your language was offered in the main menu (as option 6), while the other IVR did not offer the choice again (if you want to change your language preference, you have to speak to an agent).

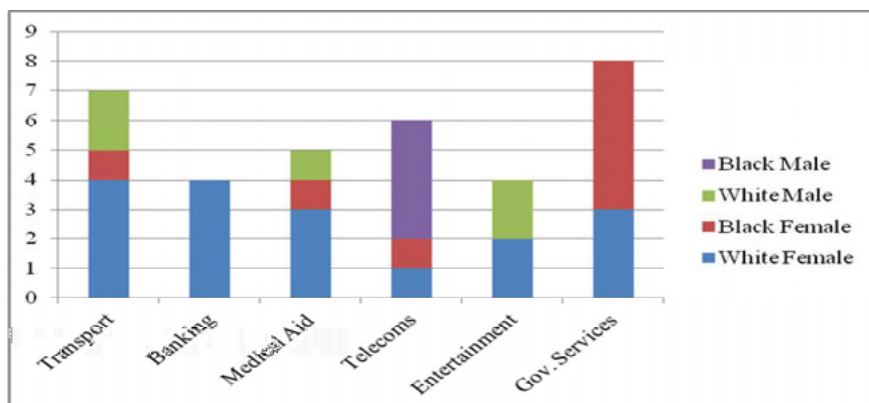
#### 4.2. Persona Design

Persona design covers many aspects relating to the “*character*” of the voice used in the IVR system, including age, gender, accent, prosody, nature of the application, brand and corporate image, etc. We investigate two aspects of the voices selected for the IVRs we researched: accent and gender.

In the IVRs we tested, there was clear preference for a female persona over a male persona, with 25 of the 34 IVRs employing a female voice and only nine employing a male voice. The female voice was preferred across all of the domains investigated, except for the telecoms domain that preferred a male voice (see Figure 1). The

entertainment domain was neutral regarding this choice, with two of the IVRs employing a male persona and the other two a female persona.

**Figure 1:** *Gender of persona per domain*



In terms of accent<sup>2</sup>, it was found that all of the IVRs employed a South African English (SAE) accent, except for one voice in the banking sector which had a (slight) British English accent.

Within the (multilingual) South African environment VUI designers typically need to consider the following issues with regard to persona design:

- Whether to use one, multilingual voice artist, or several voice artists?
- How to deal with accented speech (e.g. is there acceptance/tolerance for Black South African English accents in commercial IVRs)?
- Whether and how to ensure consistency across personas, keeping cultural differences in mind<sup>3</sup>?
- Whether the persona should be more human-like and less machine-like? (See Balentine 2007 regarding this issue in western contexts.)

The choice of one or more voice artists will usually be determined by the number of languages in which the IVR is provided. It is unlikely that one voice artist would be able to speak all of the official languages with a “standard” accent. Grouping the languages together, as mentioned in 4.1 above (see caller demographics, linguistics and external drivers), might offer the option of minimising the number of voice artists used, however, attitudes towards accented speech would need to be considered (see also Cohen *et al.* 2004). The South African VUI designer therefore needs to weigh the benefits and the drawbacks of having the same voice for all or a subset of the IVRs being developed.

To our knowledge, very little – if any – South African research exists (Ndwe 2011) on language attitudes in commercial contexts (e.g. attitudes towards using one's own language, attitudes towards accented speech, specifically BSAE, etc.). The PanSALB survey (PanSALB 2000) on language use and language interaction investigated some aspects of language attitudes in commercial contexts, including language preferred when interacting with a banking institution, or at a supermarket or shop. However, the researchers who undertook the survey could not identify definite patterns with regard to language attitudes in these domains, and as a result, proposed further investigation of this aspect.

The IVRs we investigated do not appear to have mapped the persona design to the IVR's caller demographics (language, accent or culture), opting in most cases for one language, a neutral accent and a (professional) female voice. Despite not having factual data to explain the choice of persona for the various IVRs, we assume that a female voice may have been chosen in the majority of the designs in an attempt to adhere to the "likeness principle" (i.e. the principle that end-users will more probably be attracted to a persona that is similar to themselves; Nass & Brave 2005:67). In this regard, Respondent 1 remarked that female voices are often perceived by focus groups to be more neutral, i.e. more likely to be accepted across a wider range of callers from differing cultural backgrounds.

Respondent 2 noted that his company has changed from a female to a male voice in their IVR, and although they received some negative comments right after the change, acceptance now seems to be high. This suggests that the choice of gender of the persona might not be as vital as some business owners or marketing managers might perceive. Support for the latter was found from Respondent 3; for their IVR they simply employed the call centre manager's voice, which was recorded in-house, using technology available on the IVR, without consideration given to persona design or caller demographics. The main considerations were to save costs, while still getting a somewhat "*professional-sounding voice*". Other challenges, such as call centre management and business process management, seem to far outweigh concerns over persona design.

Our conclusion here is that the gender choice for a persona is probably not a critical one, especially in light of the fact that there is limited information available on language attitudes, cultural differences and gender in commercial environments in South Africa.



### 4.3. Input Modality

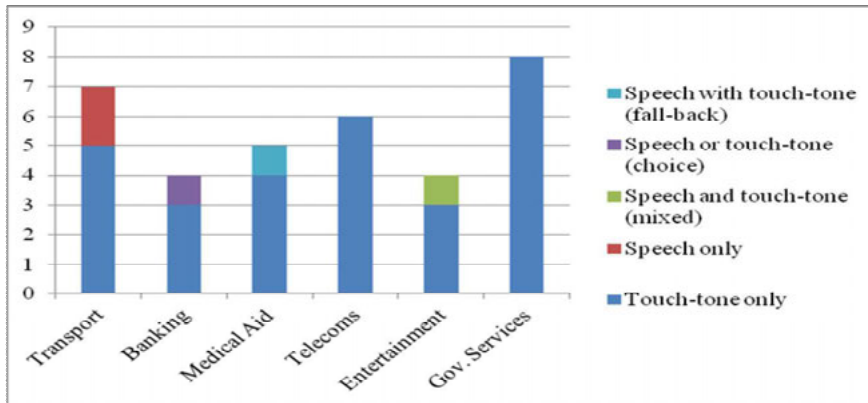
The debate on the choice of input modality in terms of speech input vs. touch-tone is a perennial topic in the design of VUIs. Much research (Lee & Lai 2005; Delogu *et al.* 1998; Franzke *et al.* 1993; Suhm *et al.* 2002) has been conducted in this area; Lee & Lai (2005) found that touch-tone is more suitable for simpler, linear tasks with a defined set of options, and speech input was more suited towards the more open domain and non-linear tasks which may have numerous options.

However, it was also noticed that user preference (towards speech) did not correlate with user performance (better with touch-tone) in terms of linear tasks (Lee & Lai 2005). Similar results have been observed by Delogu *et al.* (1998), Franzke *et al.* (1993), and Suhm *et al.* (2002), where the results indicate that the use of natural language (NL), which uses open-ended or free-form dialog, or at least some form of speech recognition, is favoured over touch-tone.

A point to note is that all of these comparison studies targeted business-typical applications, such as voicemail, call routing and banking, which are more applicable to users on the higher end of the technological-literacy spectrum in the developed world. Contrastingly, in an informal poll of over a thousand users of real-world information-access systems, almost half of the users responded that they would prefer to use a speech input modality “as little as possible” and only 8% would do so “most of the time” (Pearce & Bergelson 2008).

Similar comparison studies in the developing world have produced varying results, with some (Sharma Grover *et al.* 2009; Patel *et al.* 2009) reporting user preference for touch-tone, and others for speech (Sherwani *et al.* 2009). As observed by Barnard *et al.* (2008), in the developing world the user preference for a particular modality stems from an interplay of factors such as user ability (literacy, technology experience, user training) and application complexity (linear vs. non-linear tasks, open vs. closed domain).

In the IVRs we investigated, we find that a vast majority (29 out of 34) use touch-tone only (Figure 2). Of the remaining 5, only 2 are speech only with the other 3 featuring touch-tone as either a choice, fall-back or in a mixed input fashion. Strikingly, we notice that all 5 IVRs that use speech input are only available in English, thus from the IVRs examined that have multiple languages, none of them use speech input.

**Figure 2:** *Input modality used in IVRs*

We conjecture that there are several reasons that contribute to speech input not being used in South African multilingual IVRs:

- **Availability of commercial grade recognisers** (or language resources): To our knowledge, although there has been much progress in ASR technologies being developed for the eleven South African languages (e.g. Kamper & Niesler 2011; Mbogho & Katz 2010), commercial grade speech recognisers are only available for South African English and Afrikaans, and no such recognisers are available for the remaining nine official languages (Sharma Grover *et al.* 2011c). ASR development for the African languages is majorly slowed down by the limited number of language resources (such as language models, acoustic models, grammars and pronunciation dictionaries) available for these languages. In that context, South African English and Afrikaans have leveraged on prior work done around the world on English and Dutch respectively (Sharma Grover *et al.*, 2011c).
- **Cost of commercial grade recognisers** (or language resources): In relation to the availability of technologies is also the issue of the cost of using a commercial grade recogniser, which would currently typically be built in conjunction with an international speech vendor's technology, pushing the IVR development costs even higher. As highlighted in section 3, cost seems to be the biggest driver in developing multilingual IVRs in South Africa, all the more pushing companies towards touch-tone-based solutions, which are simpler, faster and more stable to implement. Cost implications are not only limited to development and/or licensing costs, but also to maintenance costs. As Respondent 2 pointed out, if his company would decide to implement speech, the maintenance cost would probably be enormous to regularly fine-tune acoustic models, language models, grammars, etc. for multiple languages. He also pointed out that his

company has concerns about the availability of local expertise to contract for such maintenance work.

- **Language usage:** South Africa's multicultural environment results in South African languages being spoken with several different accents. This of course increases the time, language resources and costs involved in developing a speech recogniser that caters for the various accents for a specific South African language. This is further exacerbated by the fact that code-switching (e.g. a mix between English and an African language or between two closely-related African languages) occurs frequently in South African users' speech (Roux *et al.* 2000; Finlayson & Slabbert 1997; Slabbert & Finlayson 2002). This requires a further investment in developing appropriate speech recognisers (e.g. non-native speech recognition with multiple language models), and modules, such as language identification modules, to distinguish between closely-related languages in order to address this phenomenon.
- **End-user perception:** In the developing world, user perceptions of and inexperience with ASR technologies could also rule in favour of touch-tone over speech. For example, in some previous studies (Sharma Grover *et al.* 2009) it was found that some users considered the IVR system's voice to be a real human on the other end and not automated. Whilst using touch-tone may not completely eliminate this problem, it will certainly alleviate it as there is less human-like interaction with key presses, than with speech input.

Although one might conclude from the above that emerging markets should be thankful for escaping the so-called Jetsonian culture (see International Marketing Council of South Africa 2010), where speech is often seen as the begin-all and end-all of human-computer interaction, it is not as simple as that. The advent of smart phones and phones with touch screens are likely to have a significant impact on emerging markets, and one can therefore not simply ignore ASR as a possible input modality.

## 5. CONCLUSION

In this research we aimed to get a better understanding of business and design issues related to IVRs in a multilingual, emerging market such as South Africa. Of the South African 34 IVRs investigated in this study only 9 had a multilingual offering (more than 1 language) and none of these multilingual IVRs had speech input. Speech input was available in 5 IVRs and only in English with the remaining 29 using touchtone as the input modality. In investigating business drivers for multilingual IVRs, we found that cost is probably the biggest driver for not implementing multilingual IVRs, over-shadowing many of the positive business

drivers, and the local availability of technologies and expertise are also significant hurdles for implementing ASR. Based on our limited interviews, our estimation is that South Africa will not see a dramatic increase in the number of commercial multilingual IVRs in the near future, unless (a) the customer demand increases; or (b) punitive laws and regulations force companies to have multilingual IVRs. Perhaps the example set by government bodies in implementing multilingual IVRs could in future contribute to strengthening the business case for commercial multilingual IVRs, while also generating valuable expertise on how to properly design and manage such IVRs.

Persona and gender choice for prompts seem to feature as low design priorities in the IVRs investigated. This could be attributed to the fact that there is a huge lack of business intelligence (or research findings) related to language attitudes (including attitudes towards cultural issues such as choice of persona), which makes it difficult for business managers and designers in industry to make informed choices during the design phase. This supports our contention that cost is a primary driver for multilingual IVR development in emerging markets. However, we also conclude that there are still huge opportunities for sociolinguistic research into the implications of language preferences, language attitudes and cultural differences for the development of (commercial and non-commercial) IVRs in South Africa.

In addition to questions for further research mentioned throughout the study, the following, amongst others, could also be addressed in future research – all within the context of commercial environments:

- **Human-computer interaction:** Do people from different cultures have different propensities towards using more machine-like or more human-like technologies? How does language choice influence metrics (e.g. call completion rates, time in IVR, etc.)? Should different metrics be applied to different user groups (e.g. groups with higher technology literacy than other groups)? Do people from different cultures or backgrounds have different levels of tolerance for queuing times? What are people's expectations or attitudes relating to music while waiting in a queue? Are customers content to switch between different languages (e.g. self-service vs. call centre agent)?
- **Language attitudes:** Do people need or prefer IVRs in their own languages, and if so, what are their expectations and requirements? What are their attitudes towards companies who do not offer multilingual services? How are IVRs that only offer a limited number of languages perceived? Will people use a closely-related language (e.g. isiZulu) if their own mother-tongue (e.g. isiXhosa) is not offered in an IVR? What are their attitudes towards accented speech (not only accented English)?

- **Scripting:** Are there significant differences between different language groups regarding discourse management in technology (e.g. turn-taking, greetings, apologising, error recovery strategies, time-outs and delays, etc.)? What challenge does tapering of prompts hold for different languages? What core terms are needed for broad application in IVRs (e.g. *hash, star, press, enter, zero, repeat, help*, etc.)? What are the most important issues related to style and register in IVR environments for different languages? How should product names be handled (e.g. translated or not)?

Issues such as these should of course not only be explored in the South African context, but also in other similar multilingual, emerging markets.

## ACKNOWLEDGEMENTS

We would like to express a word of gratitude to the interviewees, who made time available for interviews and/or responded to questions via email. Despite their careful formulation of answers, all (mis)interpretations and fallacies remain ours.

## ENDNOTES

1. See [www.internetworldstats.com/af/za.htm](http://www.internetworldstats.com/af/za.htm) and [www.google.com/publicdata](http://www.google.com/publicdata)
2. Perceptions of the accent of a voice artist differ from person to person. For the purposes of this research, if at least two of the authors perceived the accent as the same (e.g. as Black South African English), then it was classified as such.
3. For example, in Germany BMW had to do a product recall for their in-car navigation system that had a female voice, as German male drivers (the vast majority of BMW drivers) did not like taking directions from females (Nass & Brave 2005).

## REFERENCES

- Balentine, B. 2007. *It's Better to Be a Good Machine Than a Bad Person: Speech Recognition and Other Exotic User Interfaces at the Twilight of the Jetsonian Age*. Annapolis: ICMI Press.
- Barnard, E., M. Davel & G.B. van Huyssteen. 2010. Speech Technology for Information Access: a South African Case Study. *Proceedings of the AAAI Spring Symposium on Artificial Intelligence for Development (AI-D)*, Palo Alto, California.8-13.
- Barnard, E., M. Plauché & M. Davel. 2008. The utility of spoken dialog systems. *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*, Goa, India.13–16.  
<http://www.meraka.org.za/lwazi/publications/barnard08spokendialogue>
- Calteaux, K., A. Sharma Grover & G.B. van Huyssteen. 2012. Business drivers and design choices for multilingual IVRs: A government service delivery case study. *Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU)*. Cape Town, South Africa, May 2012.
- Cohen, M.H., J.P. Giangola & J. Balogh. 2004. *Voice User Interface Design*. Boston: Addison-Wesley.
- Department of Arts and Culture (DAC). 2003. *Implementation plan: National Language Policy Framework*. Unpublished. Pretoria: National Language Service.
- Delogu, C., A.D. Carlo, P. Rotundi & D. Sartori. 1998. Usability evaluation of IVR systems with DTMF and ASR. *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP '98)*, Sydney, Australia, November 1998.
- Finlayson, R. & S. Slabbert. 1997. 'I'll meet you halfway with language': Code-switching within a South African urban context. *Languages Choices. Conditions, Constraints and Consequences*, edited by M. Putz. Amsterdam: John Benjamins.381-421.
- Franzke, M., A.N. Marx, T.L. Roberts & G.E. Engelbeck. 1993. Is Speech Recognition Usable? An Exploration of the Usability of a Speech-Based Voice Mail Interface. *ACM SIGCHI Bulletin* 25(3):49-51.
- Gapminder. 2007/2008. Data in Gapminder World.  
<http://www.gapminder.org/data/> Accessed: 30-03-2011.

- International Marketing Council of South Africa. 2010. 2010 Human Development Index ratings.  
<http://www.imc.org.za/press-room/567-2010-human-development-index-rankings.html> Accessed: 30-03-2011.
- Internet World Stats. 2012. South Africa Internet Usage and Marketing Report.  
<http://www.internetworldstats.com/af/za.htm> Accessed: 30-03-2011.
- IT News Africa. 2009. Africa's high mobile penetration sets the stage for Internet revolution.  
<http://www.itnewsafrica.com/2009/07/africas-high-mobile-penetration-sets-the-stage-for-internet-revolution/> Accessed: 30-03-2011.
- Kamper, H. & T. Niesler. 2011. Multi-accent speech recognition of Afrikaans, Black and White varieties of South African English. *Interspeech 2011: Proceedings of 12<sup>th</sup> Annual Conference of the International Speech Communication Association*. Florence, Italy. 28-31 August 2011.
- Lee, K.M. & J. Lai. 2005. Speech Versus Touch: A Comparative Study of the Use of Speech and DTMF Keypad for Navigation. *International Journal of Human-Computer Interaction* 9(3):343-360.
- Mbogho, A. & M. Katz. 2010. The impact of accents on automatic recognition of South African English speech: a preliminary investigation. *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT)*, Bela Bela, South Africa.187-192.
- Nasfors, P. 2007. *Efficient voice information services for developing countries*. Master's thesis, Department of Information Technology. Uppsala: Uppsala University.
- Nass, C. & S. Brave. 2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Interaction Relationship*. MIT: Cambridge, USA.
- Ndwe, T. 2011. *Usability Engineering of Interactive Voice Response (IVR) Systems in Oral Users of Southern Africa*. PhD Thesis. Cape Town: University of Cape Town.
- PanSALB. 2000. *Language use and language interaction in South Africa – A national sociolinguistic report*. Unpublished research report. Pretoria: Pan South African Language Board.
- Patel, N., S. Agarwal, N. Rajput, A. Nanavati, P. Dave & T.S. Parikh. 2009. A comparative study of speech and dialed input voice interfaces in rural India. *Proceedings of the Conference on Human Factors in Computing Systems (CHI'09)*, Boston.51-54.

- Pearce, T. & M. Bergelson. 2008. Alignment index for speech self-service. Dimension Data Technical Report.  
<http://www.dimensiondata.com/NR/rdonlyres/9191A848-5F35-459F-8239-8D9D2248414E/8791/mainstreamspeechalignmentindexreport2> Accessed: 15-09-2008.
- Plauché, M. & U. Nallasamy. 2007. Speech Interfaces for Equitable Access to Information Technology. *Information Technologies and International Development (ITID) Journal* 4(1):69-86.
- Roux, J.C., E.C. Botha & J.A. du Preez. 2000. Developing a Multilingual Telephone based Information System in African Languages. *Proceedings of the Second International Language Resources and Evaluation Conference*. Athens, Greece, 31 May – 2 June 2000.  
<http://www.ast.sun.ac.za/publications/lrec2000-jcrebjadp.pdf>
- Sharma Grover, A., M. Plauché, C. Kuun, & E. Barnard. 2009. HIV health information access using spoken dialogue systems: Touchtone vs. Speech. *Proceedings of the International Conference on Information and Communications Technologies and Development (IEEE)*, Doha, Qatar.95-107.
- Sharma Grover, A., O. Stewart & D. Lubensky. 2009. Designing interactive voice response (IVR) interfaces: Localisation for low literacy users. *Proceedings of the 12th IASTED International Conference on Computers and Advanced Technology in Education (CATE)*, St. Thomas, Virgin Islands.673-680.
- Sharma Grover, A. & E. Barnard. 2011a. Comparing Two Developmental Applications of Speech Technology. *Conference on Human Language Technology for Development 2011*. Alexandria, Egypt.81-86.
- Sharma Grover, A. & E. Barnard. 2011b. The Lwazi Community Communication Service: Design and Piloting of a Voice-based Information Service. *Proceedings of the 20th WWW Conference*, Hyderabad, India.433-442.
- Sharma Grover, A., G.B. van Huyssteen & M.W. Pretorius. 2011c. The South African Human Language Technology Audit. *Language Resources and Evaluation* 45(3):271-288.
- Sherwani, J., S. Palijo, S. Mirza, T. Ahmed, N. Ali & R. Rosenfeld. 2009. Speech vs. touch-tone: Telephony interfaces for information access by low literate users. *Proceedings of the International Conference on Information and Communications Technologies and Development (IEEE)*, Doha, Qatar.447-457.
- Slabbert, S. & R. Finlayson. 2002. Code-switching in South African townships. *Language in South Africa* edited by R. Mesthrie. Cambridge: Cambridge University Press.235-257.



- Suhm, B., J. Bers, D. McCarthy, B. Freeman, D. Getty, K. Godfrey, & P. Peterson. 2002. A Comparative Study of Speech in the Call Center: Natural Language Call Routing vs. Touch-Tone Menus. *Proceedings of the Conference on Human Factors in Computing Systems (CHI'02)*, Minneapolis, Minnesota, USA. 283-290.
- Tucker, R. & K. Shalnova. 2004. The Local Language Speech Technology Initiative – localisation of TTS for voice access to information. *Proceedings of the SCALLA Conference*, Kathmandu, Nepal, January 2004.  
<http://www.elda.org/en/proj/scalla/SCALLA2004/tucker.pdf>
- United Nations Conference on Trade and Development. 2010. South-South cooperation offers new opportunities for transforming African economies.  
<http://www.unctad.org/Templates/StartPage.asp?intItemID=2871&lang=1>  
Accessed: 30-03-2011.
- Van Huyssteen, G.B., A. Sharma Grover & K. Calteaux, submitted. Offering multiple languages in interactive voice response systems.

## CHAPTER 18

# THE RELATIONSHIP BETWEEN THE AUTOMATIC ASSESSMENT OF ORAL PROFICIENCY AND OTHER INDICATORS OF FIRST YEAR STUDENTS' LINGUISTIC ABILITIES

Febe de Wet<sup>1</sup>, Thomas Niesler<sup>2</sup>, Christa van der Walt<sup>3</sup>

<sup>1</sup>*Meraka HLT Group, Council for Scientific and Industrial Research (CSIR)*

<sup>1</sup>*Stellenbosch University Centre for Speech and Language Technology*

fdwet@csir.co.za

<sup>2</sup>*Department of Electrical & Electronic Engineering, Stellenbosch University*

trn@sun.ac.za

<sup>3</sup>*Department of Curriculum Studies, Stellenbosch University*

cvdwalt@sun.ac.za

### 1. INTRODUCTION

Academic literacy proficiency is key to the success of a student at university. Currently, the large-scale assessment of language proficiency, particularly at higher education levels, is dominated by reading and writing tests because listening and speaking skills are thought to be too difficult to evaluate. The assessment of oral and aural skills is particularly challenging when large groups of students need to be considered simultaneously and when the results have to be available within a short space of time. However, to make a meaningful assessment, a balanced picture of a student's language proficiency is required, and this must include information on oral proficiency as well as listening comprehension. The application of automatic speech recognition (ASR) techniques in the automatic assessment of these skills is one of the ways in which the logistical challenges associated with testing listening and oral proficiency can be addressed.

The design and development of an automatic test for oral proficiency and listening comprehension poses a number of challenges. Firstly, the assessment of fluency above word and sentence level has to take the subjective judgements of assessors into account. In the case of extended writing assessment, attempts to obtain more objectivity include the use of rubrics or assessment schedules and multiple assessments of the same material by different assessors. The same methods can be employed for oral proficiency assessment, although it must be borne in mind that speech carries (among other attributes) the accent and gender of the testee, both of which can increase the possibility of bias. For speech fluency assessment above sentence level, which often takes the form of oral proficiency interviews, the interaction between assessor and testee may influence performance on the test,

especially when the two parties are acquainted. Hence, the possibility that subjective elements may influence the final score is much greater in oral proficiency assessment than in written tests.

Although multiple assessments of the same oral products can be introduced to obviate problems with subjectivity, they are logistically more problematic, particularly when the results must be available in a short space of time. The main difference between oral and written proficiency assessment is that, whereas we can read faster than we can write, and the assessment of writing is therefore faster, listening to speech takes as much time as the speaking itself. Even if it were feasible to record and subsequently assess a large group of students by using rubrics and an extensive group of markers, it would take much longer to conclude the assessment. Furthermore, testees may feel that listening and responding to prompts or doing an oral interview with a lecturer does not, for example, reflect their ability to teach Biology to secondary-school learners. These considerations mean that a test of oral proficiency must meet all the requirements of good assessment procedures: it must be feasible (i.e. to test large groups), reliable (to obviate problems of subjectivity) and valid (particularly in terms of face validity).

For the purposes of the project discussed in this chapter, the issue of feasibility was the main consideration for developing a test that can be administered to a large group of students in such a way that reliable results are available within a day or two. The particular setting was one in which students must be streamed into appropriate language support modules at the start of their university education. As such the test was not a final, high-stakes assessment, but the starting point for a series of more open-ended oral proficiency assessments.

Nevertheless, an automatic test can only be used as an objective measure of proficiency if the test results show a high degree of similarity with multiple human assessments of the same data. For these human assessments, the rating scales used by the evaluators must be designed very carefully to ensure reliable results. Previous attempts to design such scales showed promising results with advanced, postgraduate students (Van der Walt *et al.* 2008). In this case, however, the instrument was used with first-year students. Since students come from a variety of language backgrounds it was decided to include a wider variety of language samples, varying in linguistic and syntactic complexity. Although the students were registered for different degree programmes, they all followed a common first-year module in English Studies, which includes aspects of literature (e.g. short stories, drama and film studies) as well as aspects of language-in-use, such as the analysis of advertisements and the development of academic language proficiency. Being proficient in academic English is extremely important because even if they did not continue their study of English after the first year, the students would still

use the language in their academic subjects. Using and understanding English for cognitive, academic purposes is very important for their academic development.

In this study, all correlation values are expressed in terms of Spearman's rank correlation coefficients, which are used as a means to quantify the degree of similarity between automatic and human assessments. We report on the correlation between human and automatic ratings for a test population of first year students, as described in the preceding paragraphs. We also consider the correlation between these assessments and the performance of the same students in two written tests as well as a test of academic listening skills. In addition, we comment on the feasibility of using readability index, quantified in terms of the Flesch Reading Ease (FRE) scale, as a design criterion for test items.

The next section describes the design and implementation of the automatic test that was used in this study. Section 3 reports on the human assessment of the data and Section 4 describes the ASR system that was used to evaluate the data automatically. Section 5 introduces the other indicators of linguistic ability that our measurements will be compared to. Results are presented in Section 6. Concluding remarks and future directions are discussed in Section 7.

## **2. AUTOMATIC ORAL PROFICIENCY TESTING**

The test that was used to assess students automatically is the result of a number of piloting experiments and subsequent adjustments. The students' language background and the lecturers' expectations of their language proficiency were important considerations in the design of the test and the associated rating criteria. All the students are active bi- and multilinguals that need to use English as an academic and professional language, rather than just for everyday interpersonal communication. This orientation meant that the test needed to include context-sensitive content (i.e. educational and academic context). The rating criteria did not assess the students' proficiency in home-language speaker terms but rather in terms of intelligibility and comprehensibility. Although accent and grammatical correctness played a role in the assessment criteria, they were only taken into account when comprehensibility was affected.

Our initial aim was to develop a test of oral and listening proficiency in English for postgraduate students doing a certification course to become secondary school teachers. The first version of the test included seven sections and was implemented as a telephone-based spoken dialogue system (SDS) (Van der Walt *et al.* 2008). However, the students took much longer to complete the test than was anticipated. Subsequent versions of the test were therefore limited to a reading and a repeating (elicited imitation) task. The purpose of the reading task is to test comprehension of the instructions and the ability to read fluently using appropriate intonation and

pronunciation, while the elicited imitation task aims to determine the extent to which students can grasp the meaning of what they hear and repeat and/or rephrase what they heard.

Experiments involving different groups of post-graduate students consistently showed that the students did not find the reading task challenging, and obtained very high scores. This left little room for discrimination between different levels of proficiency. The elicited imitation task, on the other hand, resulted in a wider range of scores and showed more potential as an indicator of oral proficiency (De Wet *et al.* 2009, De Wet *et al.* 2010).

This kind of task is controversial and seems to have originated in the field of language therapy with a view to predicting spontaneous speech production (Fujiki and Brinton 1987). The prompts consist of sentences that are just longer than can be accommodated with ease by the students' working memory. Graham *et al.* (2008:1604) explain the process as follows:

Since short-term or working memory is limited, the retention of a representation there is, by most accounts, dependent upon the number of units being processed. As the length of utterances becomes greater, it necessitates the chunking of information ... It is believed that language competence is what facilitates this chunking process.

Since pilot versions of the test showed that this kind of test item discriminated more consistently among advanced users of English, it seemed a promising direction on which to focus in subsequent versions of our automatic assessment system. Moreover, it is an important skill for higher education students to listen and make notes in lectures and therefore the ability to retain and process complex sentences, albeit in this case in writing. This is an important motivation for assessment by elicited imitation. The current version of the test consists of a SDS, running on a desktop PC and administered in a multi-media computer laboratory using headsets with directional, noise cancelling microphones. During the test, students were prompted to read sentences from a test sheet as well as to repeat utterances produced by the SDS. On average, the students took around seven minutes to complete the test.

## 2.1. Prompt Design

In an effort to test the effect of varying sentence difficulty on reading ability and repetition accuracy, it was necessary to find a consistent means of calculating both the complexity and length of sentences. The level of sentence difficulty for the test was therefore linked to the FRE scale. The FRE scale is a standard means to quantify textual sentence complexity and provides an indication of how easy or difficult a given text is to read: higher FRE scores are associated with easier texts

and lower scores with more difficult texts. We chose to use the FRE scale because it is readily available and easy to use, e.g. it is included as a standard tool in most word processing packages. Such readability scores are often criticised, because they are generated at sentence level and do not take the overall structure of a text into account (Shehadeh and Strother 1994). However, since the test was designed with a focus on the sentence level, the FRE scale was regarded as suitable for our purposes.

For the reading task, the readability scores ranged from 28.5 to 83.0 (average = 52.3). The prompts for the elicited imitation task were divided into two sets: a fairly easy (*Repeat A*) and more challenging (*Repeat B*) set. The readability scores ranged from 65.7 to 85.2 (average = 70.5) and from 46.6 to 57.7 (average = 50.8) in *Repeat A* and *Repeat B*, respectively. Since previous attempts to use elicited imitation indicated the importance of context when using this technique (Fujiki and Brinton 1987:302) and since the purpose here was to assess the oral proficiency of multilingual, higher education students who use English mainly for educational purposes, the vocabulary of both tasks was controlled by focusing on educational settings with which participants would be familiar, for example:

- Read: *First year students find that they lack academic skills.* (FRE = 71.7)
- Repeat A: *I don't see useful teaching techniques in the schools.* (FRE = 85.2)
- Repeat B: *Teachers often resist change and don't want to see new methods, unfortunately.* (FRE = 54.2)

In total, there were fifteen possible sentences in the reading task, seven sentences in *Repeat A* and eight in *Repeat B*. During the test, the SDS randomly selected six sentences to be read: three easier and three more challenging repeat prompts from *Repeat A* and *Repeat B* respectively. Each student was therefore prompted to read and repeat six utterances.

## 2.2. Test Population

The automatic test was taken by 58 first-year undergraduate students at Stellenbosch University. The students were divided into three groups in terms of their overall performance in their first year English Language and Literature module. The first group contained those students achieving an average mark between 40 and 49%, the second between 50 and 59%, and the third between 60 and 75%.

A mark of 40% is the lowest possible, while 75% represents the top mark among the 58 students. From these three groups, random selections were made to obtain a spread of participants from low to high scoring individuals. In this way, the number of participants per language proficiency group could be kept equal across

groups to facilitate comparison between oral proficiency scores and written test scores achieved in their English module.

### 2.3. Test Administration

Oral instructions were given to the students before the test. They also completed an audio test to verify playback and recording volume. We included the audio test because, in previous studies, we found that some students spoke too loudly (causing clipping in the audio file) while others spoke too softly to perform meaningful ASR.

In addition to the instructions given by the SDS, a printed copy of the test instructions was provided. The students were allowed to read through the instructions before taking the test. No staff was present while the students were taking the test.

## 3. HUMAN ASSESSMENT

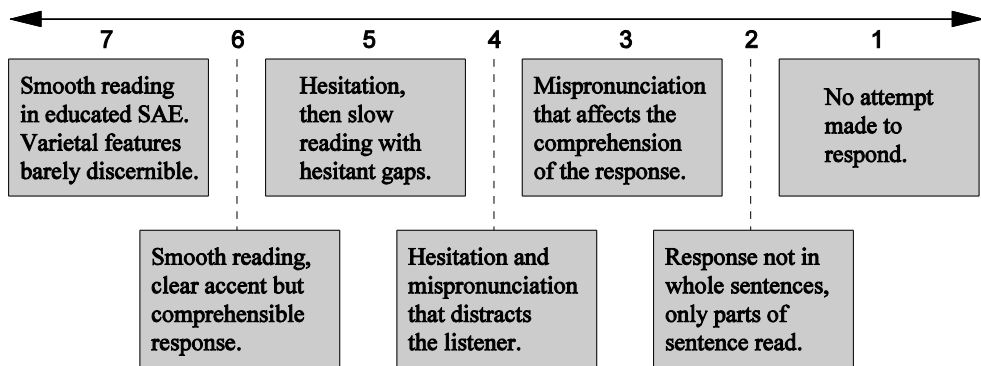
Six teachers of English as a second or foreign language were asked to rate the students' responses. Each teacher also rated at least three students twice. The intra-rater reliability for the majority of the teachers was above 0.9. Inter-rater agreement was determined in terms of two-way, intra-class correlation coefficients (ICCs). The ICC values indicated that the evaluations of two of the teachers differed substantially from those of the other four ( $ICC < 0.2$ ). The ratings of these two teachers were therefore not taken into account during the rest of the study. The inter-rater agreement for the four remaining teachers varied between 0.87 and 0.88.

The rating scales for the reading and elicited imitation tasks are illustrated in Figures 1 and 2 respectively. These scales were developed by taking the results of a number of previous rating experiments into consideration (De Wet *et al.* 2009). For the reading task, the aim was to assess the students' ability to read without hesitation and with pronunciation and phrasing that closely approximates what would be regarded as educated South African English in which an L2 accent may be discernible. This scale allows for active bi- and multilinguals who are not necessarily attempting to sound like home language speakers of English but who are competent users of the language for academic purposes.

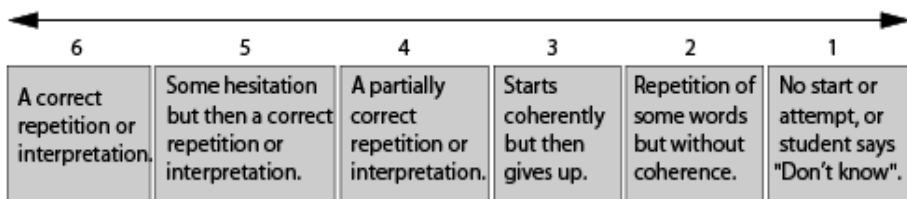
The scale for the elicited imitation task was designed to assess sentences in terms of the degree of hesitation as well as the accuracy of *repetition* or *interpretation*, which meant that students could still be given the highest score even if they did not use the exact words or the exact word order of the prompt. In language processing terms an accurate interpretation would mean that the working memory is able to make meaning of the incoming message by repeating its essence, a phenomenon

that was observed in earlier versions of the test. For example the sentence, “*Teachers often resist change and don't want to see new methods, unfortunately*” could be reinterpreted as “*Unfortunately teachers don't want to change or use new methods*” and still be given the highest score on the human ratings. One could even argue that an accurate interpretation shows language proficiency at a more advanced level than mere repetition of the original prompt.

**Figure 1:** Scale used by humans to rate read prompts.



**Figure 2:** Scale used by humans to rate repeated prompts.



The raters were not provided with the numerical values indicated in the figures. These were used only to quantify the ratings for subsequent correlation with machine scores.



#### 4. AUTOMATIC ASSESSMENT

The output of an ASR system can be used in various ways to extract quantitative features from speech signals. Various techniques to derive these so-called *machine scores* from speech data have been reported on in the literature (Speech Communication 2000, Speech Communication 2009). In previous studies, we consistently found low correlation between human ratings and what are known as posterior scores (De Wet *et al.* 2009). This investigation will therefore be restricted to scores derived from segmentation information and from repeat accuracy, because, to date, these have shown the highest correlation with human ratings (De Wet 2010).

##### 4.1. The ASR System

The Hidden Markov Model Toolkit (HTK) version 3.4 was used for ASR (Young *et al.* 2006). The hidden Markov models (HMMs) used by the speech recogniser were trained on approximately 6 hours of telephone quality speech from English mother-tongue speakers. An equal number of male and female speakers were included in the speaker population. This data is part of the African Speech Technology corpus, and consists of phonetically and orthographically annotated speech gathered over South African fixed as well as mobile telephone networks (Roux *et al.* 2004).

The test data was bandlimited during feature extraction to match the frequency range of the training data. Features were extracted from 25 ms overlapping frames of acoustic data and subsequent frames were extracted every 10 ms. Each acoustic data frame was encoded as 12 Mel-Frequency Cepstral Coefficients (MFCCs) and an energy feature (C0). Cepstral mean normalisation was applied at utterance level. The first and second order derivatives were extracted from the static coefficients and appended to the feature vector.

Triphone HMMs were obtained by means of decision-tree state clustering and embedded Baum-Welsh re-estimation. The final set of triphone HMMs consisted of 4797 tied states based on a set of 52 phones, and a maximum of 8 Gaussian mixtures per HMM state.

Finite State Grammars (FSGs) were used for the automatic recognition of the reading task. It is expected that the students, who generally have good English reading skills, would make very few errors while reading prompts from a test sheet. Hence the use of a strict finite state grammar (FSG) is an appropriate recognition method for this task. For each prompt in the reading task an FSG was created allowing the desired utterance, as well as "I don't know" or simply "don't know". The branch allowing the desired utterance expects all words to be present

in the correct order, but allows inserted silence, noise and filled pauses. These prompt-specific grammars were defined using extended Backus-Naur form (EBNF) notation and were parsed to lattice files that were used during recognition.

Unigram language models (LMs) were used for the automatic recognition of the repeated prompts. For this task, provision must be made for missing words and changes in word order. A separate unigram language model (LM) was therefore created for each prompt of the elicited imitation task. Each LM consisted of an unweighted word loop, with word-to-word transitions between the words in the prompt all having an equal probability. Silence, noise and filled pauses were allowed between words.

Since the word insertion penalty and language model scale factors were optimised on a development set in previous experiments (De Wet *et al.* 2009), all data collected in this study was available for use as a test set.

## 4.2. Segmentation-Based Scores

The scores based on segmentation information focus on the *temporal* features of speech, rather than on its acoustic characteristics, and are calculated from phone level alignments. A distinction is made between the *speech phones* (those forming part of words) and *non-speech phones* (those forming part of silence or noise) in each utterance. In a previous study, four segmentation-based scores were investigated: rate of speech, articulation rate, phonation/time ratio and segment duration scores. The highest correlation between a segmentation based score and the human ratings of the same data was observed for rate of speech (De Wet 2010), and the scope of this study will therefore be restricted to this measure.

### 4.2.1 Rate of Speech

The *Rate of Speech* (ROS) of an utterance is defined in Cucchiari *et al.* (2000) as the number of speech phones per second, calculated using the number of speech phones in the utterance  $M_{\text{Speech}}$ , and the total duration of the utterance  $T_{\text{Total}}$ , in seconds:

$$ROS = \frac{M_{\text{Speech}}}{T_{\text{Total}}}$$

Any silences leading or trailing the utterance are ignored when determining the total duration.

### 4.3. Scores Derived from Repeat Accuracy

Speech recognition accuracy can also be used as a score for automatic assessment. Previous experiments have shown that most university level students are able to achieve a perfect reading accuracy for the majority of the prompts in the exercise. Reading accuracy is therefore not useful in scoring proficiency automatically. Repeat accuracy, on the other hand, is more variable and two closely-related alternatives were considered in this study: ASR Accuracy and ASR Correct.

#### 4.3.1 ASR Accuracy

The score *ASR Accuracy* ( $Acc_{ASR}$ ) is calculated using the HTK tool *HResults*, which uses a dynamic programming-based string alignment procedure to align the recogniser output with the reference transcription (Young *et al.* 2006). It counts the number of correctly aligned words ( $H$ ), the number of insertions ( $I$ ), and the number of words in the reference transcription ( $W$ ).

The score is then calculated as:

$$Acc_{ASR} = \frac{H - I}{W} \times 100\%$$

Note that this score is penalised by insertions. When the number of insertions exceeds the number of correctly recognised words, the score is negative.

#### 4.3.2 ASR Correct

The score *ASR Correct* ( $Cor_{ASR}$ ) indicates the percentage of reference transcription words present in the recogniser output (Young *et al.* 2006).

$$Cor_{ASR} = \frac{H}{W} \times 100\%$$

In contrast to  $Acc_{ASR}$ , this score does not take insertions into account, but simply reflects the percentage of correctly-aligned words.

## 5. INDICATORS OF LINGUISTIC ABILITY

The two written tests that were chosen as indicators of linguistic ability are the so-called *Test of Academic Literacy Levels* and the *Early Assessment* test. Students' performance on the *Academic Listening Test* was also taken into consideration.

### **5.1. Test of Academic Literacy Levels**

The Test of Academic Literacy Levels (TALL) is a multiple choice test of comprehension, academic vocabulary, inference, coherence and register (Van Dyk and Weideman 2004). Its purpose is to assess the existing academic literacy of incoming students with a view to streaming them into appropriate language support modules. All first-year students are expected to complete this test.

### **5.2. Early Assessment Test**

The Early Assessment test (EA) is a university-wide measure of how first-year students perform in their various modules after six weeks in the first semester of the first year of undergraduate study. The purpose is to identify students who are at risk of not passing and to provide appropriate academic support. The EA score can include a number of assessments, depending on the structure of particular modules. In the case of the students who participated in this study, the EA consisted of an academic essay.

### **5.3. Academic Listening Test**

The Academic Listening Test (ALT) was developed at Stellenbosch University with the express purpose of appropriately assessing the listening proficiency of first year students, using academic material and an academic context (Marais and van Dyk 2010). Students completed the computer-based test by answering multiple-choice questions that elicited responses in four tasks:

- Students were required to structure information;
- Students were required to watch a video-recorded lecture and subsequently answer questions on, for example, its main and supporting idea;
- Students were required to watch a video-recorded discussion by two students and subsequently answer questions on, for example, the represented attitudes;
- Students were required to listen to a video-recorded lecture and subsequently fill in words that had been omitted from a transcript.

The ALT consisted of multiple-choice questions throughout.

6. RESULTS

The previous sections have described how language proficiency can be assessed by means of written tests, by means of oral assessments with human evaluators, and by means of automatically-derived oral proficiency indicators. We will now investigate how well these various approaches relate to one another.

6.1. Relationship between Human and Automatic Oral Proficiency Assessments

The correlations between the machine scores defined in Section 4 and the ratings given by the English teachers are shown in Table 1. All the correlations in the table are statistically significant ( $p$ -values  $< 0.05$ ). The results for the elicited imitation task are shown separately for *Repeat A* and *Repeat B*, as well as for the exercise as a whole (*Repeat*). For each utterance, the human rating was taken to be the average of the individual scores given by the four judges, as described in Section 3.

Table 1: Correlation between human ratings and automatically-derived scores

	ROS	ACC <sub>ASR</sub>	COR <sub>ASR</sub>
Read	0.40	-	-
Repeat A	0.47	0.42	0.42
Repeat B	0.65	0.49	0.84
Repeat	0.55	0.36	0.67

From Table 1 we see that the correlation between *ROS* and the human ratings for the reading task is low. Other researchers have shown *ROS* for read speech correlates very strongly with human ratings of fluency (Cucchiarini *et al.* 2000). However, a number of our experiments have shown that *ROS* no longer correlates well with human ratings when the test subjects are very proficient speakers (De Wet *et al.* 2009).

The results in Table 1 indicate that the correlation between *COR<sub>ASR</sub>* and the human ratings of repeat accuracy (as defined in Figure 2) are much higher than the corresponding values for *ACC<sub>ASR</sub>*. *ROS* is also better-correlated with the human ratings of repeat accuracy than *ACC<sub>ASR</sub>*.

In general, the correlations associated with *Repeat B* are higher than those measured for *Repeat A*. This observation is attributed to the much wider range of human ratings and corresponding machine scores in *Repeat B* than in *Repeat A*. Higher

correlations are usually observed for scores spanning wider ranges of the assessment scale than for those limited to a small interval. The correlations between the human and automatic scores are also in the same range as in previous studies, even those where a bigger group of human raters were involved in the assessment. These trends and results are in good agreement with those from previous studies involving post-graduate students, which indicates some consistency of the test over different test populations (De Wet *et al.* 2009, De Wet *et al.* 2010).

## 6.2. Relationship between Written and Oral Language Proficiency Assessments

When considering the language proficiency assessments described in Section 5 in isolation, we find that they are poor indicators of one another. For example, the correlation between the results of the TALL and the ALT is 0.52, while it is just 0.33 for the TALL and the EA ( $p > 0.05$  in both cases). These low correlations indicate that the three tests probably assess different aspects of linguistic ability.

When we consider how well the written tests mirror the results of human or automatic oral assessments, the picture described by Table 2 emerges. Only correlations that are statistically significant are shown ( $p < 0.05$ ).

**Table 2:** *Correlations between oral proficiency assessments (human and machine) and other indicators of linguistic ability*

	TALL	EA	ALT
Human raters (Read)	0.48	0.42	0.47
ROS (Read)	-	-	0.44
Human raters (Repeat B)	0.55	-	0.49
ROS (Repeat B)	0.39	-	0.41
Cor <sub>ASR</sub> (Repeat B)	0.35	-	0.37

The correlations between *ROS* and *Cor<sub>ASR</sub>* for *Repeat B* and the TALL results are lower than the correlation between the human oral proficiency assessments for *Repeat B* and TALL (0.39 and 0.35 as opposed to 0.55). This indicates that the two automatically derived measures are poor indicators of written proficiency.

From Table 1 we recall that the correlation between *ROS* and the human assessments for *Repeat B* was 0.65 ( $p < 0.05$ ). The corresponding value for *Cor<sub>ASR</sub>* was 0.84 ( $p < 0.05$ ). These figures are considerably higher than the correlation between the TALL or ALT results and the human assessments for *Repeat B* (0.55

and 0.49 respectively). This indicates that the automatically-derived measures are better indicators of the human assessments of oral proficiency than either of the written tests.

For the EA test, the correlation with the human assessments for read speech are low and significant (0.42). This indicates that the human ratings cannot be used to predict performance on the early assessment tests.

Overall, the results in Table 2 seem to indicate that the oral assessment is not a good indicator of performance in the EA tests. However, neither is the TALL, despite also being a written test. It should be borne in mind, though, that the TALL is in multiple choice format while the EA test is in the form of an essay. Although a positive correlation was found between the TALL scores and a short, open-ended writing piece that was included in early versions of TALL (Van Dyk and Weideman 2004), this is not the case in our study.

6.3. Relationship between Human Assessments and the FRE Scale

Table 3 shows the correlation between the per-utterance average of the human scores and the difficulty of the corresponding utterance on the FRE scale. As in Table 1, the results for the elicited imitation task are shown for *Repeat A* and *Repeat B* separately as well as for the whole exercise (*Repeat*). The p-values associated with the correlations are given in the third column of the table.

**Table 3:** *Correlation between human ratings (judge average) and utterance difficulty according to the FRE scale*

	Correlation	p-value
Read	0.37	0.17
Repeat A	0.49	0.26
Repeat B	0.57	0.14
Repeat	0.86	<0.00

The results in the first row of Table 3 reject the hypothesis that there is a correlation between the FRE scale level and the human ratings of the read prompts. The same trend is observed if *Repeat A* and *Repeat B* are considered separately. However, for a combination of the easy and more challenging tasks, there is a high and significant correlation between the FRE scale levels and the human ratings of repeat accuracy. This seems to indicate that FRE scale values can be used as a design criterion, provided that the exercises include utterances with varying levels of difficulty - as

is, indeed, required by test designs that attempt to assess overall speaking proficiency (Graham *et al.* 2008).

## 7. DISCUSSION AND CONCLUSIONS

Our journey through the various versions of the ASR assessment has led to questions on a number of levels and in a variety of fields.

Firstly, the use of ASR to assess oral proficiency shows the complex interplay between psycholinguistic processing, like the role of working memory, and sociolinguistic factors, like the role of context, in the oral production of language. Secondly, and at the same time, it requires of researchers from very different disciplinary backgrounds to collaborate on an assessment that has consequences for students in terms of the level and kind of academic support that they need. Thirdly, the original intention of this project - to investigate the possibility of ASR as a valid, reliable and feasible instrument - led to comparisons with other kinds of assessment. Although the results discussed above provide information about the use of ASR as a measure of oral proficiency, they also say much about the kinds of assessment conducted with first year students and the lack of correlation among these tests.

In earlier versions of the test it was not possible to find a positive correlation between ASR scores and ratings of short, open-ended oral proficiency tasks, which may be an indication of the degree to which these tasks differ in context and content (Müller 2010). Although one could argue that these assessments provide an overall picture of proficiency in different areas, it remains the case that there is no easy way for lecturers in language support to assess and predict performance for an overarching construct such as 'academic language proficiency'.

The ASR test shows positive correlations with human ratings of the speech sample, with the highest rating for  $Cor_{ASR}$  on the elicited imitation tasks. The human rating scales on these tasks required that they award highest score for a correct repetition or interpretation, which meant that students could still be given the highest score even if they did not use the exact words in the exact order. The human ratings therefore allowed for a meaning-making process that would repeat the essence of the original prompt. For an ASR system this is quite a challenge and yet  $Cor_{ASR}$  appears to measure this adequately and produced the highest correlation between the human ratings and machine scores. The fact that this measure does not penalise for insertions or require a 100% accurate repetition of words in the same order as that of the original (as in the  $Acc_{ASR}$  measure) means that an appropriate degree of flexibility is built into the system, resulting in a high correlation with human ratings.



The second highest correlation with human ratings was with the *ROS* measure. At first glance, and based on other studies, the benefit of *ROS* as a measure of oral fluency is that it can be measured by simply providing students with texts to read (Cucchiaroni *et al.* 2000). However, in this study, the *ROS* scores on the elicited imitation task correlated far better with human ratings. This result also confirms the value of using readability scores to develop test items for elicited imitation tasks.

Finally, our experiments confirm the widely-held belief that written tests are a poor indicator of oral proficiency. Furthermore, we demonstrate that measures derived automatically from a recorded speech signal are better indicators of oral proficiency than the written tests, because they show substantially higher agreement with the opinions of human judges. This indicates that an automatic oral proficiency assessment system has a clear additional role to play in the evaluation of language skills.

Future research will focus on a number of issues. Firstly, we will consider the implementation of an ASR system trained on a representative variety of South African English accents for the calculation of the automatic proficiency indicators. By experimental evaluation, we will determine whether more advanced ASR benefits the accuracy of the automatic assessment system. Secondly, we would like to automatically include synonyms in our unigram language models as used to determine the repeat accuracy. This would allow the system to be more flexible and not to penalise responses that are semantically equivalent to the prompt. Finally, we would like to expand the number of test items while maintaining a wide variety of difficulty levels. This we believe can be achieved in a semi-automatic manner with the help of the FRE scale.

## ACKNOWLEDGEMENTS

This research was supported by an NRF Focus Area Grant for research on *English Language Teaching in Multilingual Settings* as well as NRF grants TTK2007041000010 and GUN2072874 and the *Development of Resources for Intelligent Computer-Assisted Language Learning* project sponsored by the NHN.

## REFERENCES

- Cucchiaroni, C., H. Strik & L. Boves. 2000. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication* 30:109-119.
- De Wet, F., C. van der Walt & T.R. Niesler. 2009. Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication* 51(10):864-874.
- De Wet, F., P.F. de V. Müller, C. van der Walt & T.R. Niesler. 2010. *Using segmentation and accuracy-based scores to automatically assess the oral proficiency of proficient L2 speakers*. Proceedings of the 21<sup>st</sup> Annual Symposium of the Pattern Recognition Association of South Africa. Stellenbosch, South Africa, 2010.
- Fujiki, M. & B. Brinton. 1987. Elicited imitation revisited: A comparison with spontaneous language production. *Language, speech and hearing services in schools* 18:310-311.
- Graham, C.R., D. Lonsdale, C. Kennington, A. Johnson & J. McGhee. 2008. *Elicited imitation as an oral proficiency measure with ASR scoring*. Proceedings of the 6<sup>th</sup> Conference on Language Resources and Evaluation (LREC). Marrakech, Morocco.1604-1610.
- Marais, F.C. & T.J. van Dyk. 2010. Putting listening to the test: An aid to decision-making in language placement. *Per Linguam* 26(2):34-51.
- Müller, P. F. de V. 2010. *Automatic Oral Proficiency Assessment of Second Language Speakers of South African English*. Master's Thesis. Stellenbosch: Stellenbosch University.
- Roux, J.C., P.H. Louw & T.R. Niesler. 2004. The African Speech Technology Project: An Assessment. *Proceedings of LREC*. Lisbon, Portugal. Vol.1:93-96.
- Shehadeh, C.M.H. & J.B. Strother. 1994. *The use of computerized readability scores: Bane or blessing?* Proceedings of the Annual Conference of the Society for Technical Communication.41:225.
- Speech Communication. Various Authors. 2000. *Special Issue on Language Learning*. *Speech Communication* 30(2-3).
- Speech Communication. Various Authors. 2009. *Special Issue on Spoken Language Technology for Education*. *Speech Communication* 51(10):831-1038.

- Van der Walt, C., F. de Wet & T.R. Niesler. 2008. Oral proficiency assessment: the use of automatic speech recognition systems. *Southern African Linguistics and Applied Language Studies* 26:135-146.
- Van Dyk, T.J. & A.J. Weideman. 2004. Switching constructs: on the selection of an appropriate blueprint for academic literacy assessment. *Journal for language teaching* 38(1):1-13.
- Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev & P. Woodland. 2006. *The HTK book, version 3.4*. Cambridge: Cambridge University Engineering Department.

## CHAPTER 19

### CONCEPTS FOR DIFFERENT TYPES OF INFORMATION TOOLS

Henning J. Bergenholtz

*Centre for Lexicography, Aarhus University, Aarhus, Denmark*  
*Department of Afrikaans and Dutch, Stellenbosch University, South Africa*  
hb@asb.dk

#### 1. LANGUAGE TOOLS AND INFORMATION TOOLS

It would be appropriate to honour our esteemed colleague by focusing on one of the main areas of his research, namely speech recognition and speech synthesis. In this field, as in my own (lexicography), empirical analyses are undertaken and concrete conceptual proposals for tools are formulated; in the case of Justus C. Roux, these are speech technology tools and in my case they are information tools. Besides Roux's contributions (e.g. Roux 2011) to lexicography and speech technology, however, this is where the commonality ends.

As regards the research undertaken by our guest of honour, it may not be much of an issue to describe his fields of research as applied linguistics or something similar. For lexicography the situation is different – not only for many theoretical reasons, but especially because the research objectives and the methodological tradition of linguistics have had a very adverse influence on the theoretical and practical lexicography of the past 40 years. This negative finding will be elucidated and discussed first, but will then be presented with constructive suggestions with reference to concrete internet dictionaries. Although these are all electronic dictionaries, exactly the same arguments are applicable to printed dictionaries (some of the electronic dictionaries mentioned have also appeared in print).

#### 2. LEXICOGRAPHY IS PART OF INFORMATION SCIENCE

Normally, if in interdisciplinary research and the elaboration of concrete speech or information tools one holds the classification of disciplines to be an important matter and devotes too much attention to it, this is not very productive. It is productive only if the result of a specific classification is that the research and the practical results of this research are affected negatively. This actually happened when the lexicographic territory was occupied by linguistics, as will be shown below.

First, however, it will be demonstrated briefly why the compilation of dictionaries cannot be a linguistic discipline. Let us take the example of the four Danish music dictionaries, the concept of which will be presented later on. The following collaborated on these dictionaries: 1) a musicologist, 2) a computer expert and 3) a lexicographer. The latter worked out the basic outlines of the concept in agreement and after discussions with the two other collaborators, but he did not write a single dictionary entry (of course not, since he is not an expert on music). The musicologist, who entered all the data into the respective fields of the music database all by herself, is not a linguist. The fact that she is able to read Danish, German, English, French and Italian manuals and to write Danish does not make her a trained linguist at all. The musicologist did not use a text corpus during the work either, of course, as linguists demand for each and every lexicographic task these days, since the database contains no collocation fields; neither should it have such fields, since no communicative function of the music dictionaries which were compiled requires such information. However, many textbooks and music dictionaries – often called music lexicons – were used to complete missing and uncertain subject knowledge.

The reason why many people assume that lexicography is a linguistic discipline is related to the fact that linguists act as subject experts in the compilation of general communication (receptive, text production and translation) dictionaries. Over time, such linguists then also become lexicographers and also occupy themselves theoretically with lexicographic problems, frequently – or rather, usually – in the role of systematising describers of existing general dictionaries. Of course, this would not have been a problem per se if it were not associated with substantial theoretical and practical disadvantages to practical and theoretical lexicography.

The problem field comes to the fore most clearly in the British lexicographic tradition, which maintains that lexicographic theories are totally unnecessary or even impossible and acknowledges only the existence of linguistic theories, which are held to be very useful to any and all types of lexicographic practice:

This is not a book about ‘theoretical lexicography’ – for the very good reason that we do **not** believe that such a thing **exists**. But that is not to say that we pay no attention to theoretical issues. Far from it. There is an enormous body of linguistic theory which has the potential to help lexicographers to do their jobs more effectively and with greater confidence. (Atkins & Rundell 2008:4)

This rejection of theory is common among British lexicographers, or lexicographers working in the British tradition, as found in a book by a French linguist:

There are theories of language, there may be theories of lexicology, but there is no theory of lexicography. (Bejoint 2010:161)

I would call them linguists who are interested in dictionaries.

The question arises which linguistic theories the musicologist – or the computer specialist, or the lexicographer – should have applied when compiling the above-mentioned music dictionaries. As it is, the latter has in many contributions occupied himself with topics which are in his view lexicographically relevant, but not linguistically; see e.g. Bergenholtz (1996, 1998), Almind and Bergenholtz (2000), Bergenholtz and Tarp (2002, 2003), Bergenholtz and Johnsen (2005, 2007), Bergenholtz and Gouws (2007).

In large parts of continental European lexicography, especially in the Nordic and German lexicography, lexicography is regarded without reserve as a scientific and even theoretical discipline. There are also professorships specifically for lexicography and many associated positions at the respective universities. The most striking development of theory is taking place in the German-speaking regions. I will explain below why I am nevertheless not convinced that these theories will be of great future benefit. It will become clear that while the contributors do argue theoretically, they must in practice be regarded as being basically linguists who, although they wave the banners of 'user' and 'user-friendly', still essentially think and argue like linguists and not as representatives of a field that focuses mainly on information tools for non-linguists and only in exceptional cases on information tools for specialist linguists and students of linguistics with cognitive data, i.e. data relating to knowledge.

### **3. NEW CONTRIBUTIONS TO INTERNET LEXICOGRAPHY**

I would like to illustrate the critical thesis with some quotations from one of the latest editions of an international lexicographic journal, namely *Lexikographica* 26, 2010. I use only quotations from two of the total of 18 contributions to internet lexicography in the relevant edition of the journal; the other 16 contributions basically follow the same trend (all quotations translated from German):

The more information is offered, the more difficult it becomes for the user to find exactly the information he needs. (Haß & Schmitz 2010:4)

It is true that some internet dictionaries are slower to access than printed dictionaries; for examples with access times see Bergenholtz (2009) and Bergenholtz and Gouws (2010b). However, this was not because those with slower access contained larger amounts of data, but mainly because the technical means provided were not integrated. What is more important is the fact that the dictionaries in question were polyfunctional dictionaries with up to several web pages per entry. The term 'polyfunctional' refers to the type of dictionary that is typical of printed dictionaries: a single dictionary is compiled with communicative as well as cognitive functions and targets a very broad group of users. This may be justified if one does not print a separate reception dictionary, a separate text

production dictionary for mother-tongue speakers, a separate production dictionary for foreign-language speakers, a documentation dictionary for linguists and others specifically interested in language etc., as it is simply not feasible. If one has a lexicographic database, however, one can with little effort compile as many polyfunctional dictionaries as one deems necessary. This is the only way to avoid the information overload Haß and Schmitz (2010) identify as a threat, but for which they fail to offer any of the known and easily implementable solutions.

The more different functions an online dictionary fulfils, the more prior knowledge the user needs and the clearer he must formulate his questions. (Haß & Schmitz 2010:4)

This is in essence a further example of the same traditional view that dictionaries must be polyfunctional, like most printed dictionaries to date. If a concrete online dictionary, like some printed dictionaries, is designed as an aid to the reading of texts – for example Leth (1800), a pure reception dictionary for young people who want to read religious texts – it should only explain the meaning of a word or phrase and nothing more. If a user of the Danish music dictionary (Meaning of Music Terms 2011) looks up the term *zarzuela*, which he read in the booklet of a CD with Spanish music and does not understand, he finds the answer to his question, and nothing more than that:

*zarzuela*

**Kort forklaring**

en spansk form for syngespil med talt dialog

*zarzuela*

**Short explanation**

a Spanish form of ballad opera with spoken dialogue

No prior knowledge is needed to find this entry; one needs merely to be able to select between four possibilities, one of which is the button "understand a musical expression". A good tool is one that is set up in such a way that it is easy to use without any user instructions. This also applies to information tools. This music database contains many more data types, but in this case only the data in one data field is retrieved. There is no information overload, and a long learning phase is not required either. This is also true, in fact especially true, if one builds up a multilingual and particularly a technical database and uses it as a basis for a single dictionary instead of for many dictionaries:

Thus the traditional distinction between different types of dictionaries (e.g. mono- or multilingual, onomasiological or semasiological, general or specialist, etymological etc) is replaced with comprehensive documentation. (Haß & Schmitz 2010:3)

It is striking that all data in a database are equated to the data in a concrete dictionary. This is not only old-fashioned in the sense of the databases of the eighties of the previous century; it is also much too linguistic in the sense of a philological science that documentation should be the purpose of such a dictionary

or of an entire database. No doubt there is a need for scientific documentation dictionaries. But if they are to be good tools, not all communication dictionaries have to document; they should instruct, i.e. they should give clear answers to a question that arose in a communication situation. At the Centre for Lexicography of the University of Aarhus, an accounting database was compiled in collaboration with the university in Valladolid. This database contains entries in the specialist languages US English, UK English, English according to the international standards (IAS/IFRS), Danish, Danish according to the IAS/IFRS, Spanish as well as Spanish according to the IAS/IFRS. This database is the basis for several printed dictionaries such as Nielsen *et al.* (2007) or Fuertes Olivera *et al.* (2010), but also for no less than 22 different monofunctional internet dictionaries. A user who experiences a specific text production problem when writing an English text does not get entries from all fields of the database, but only the relevant entries with grammatical information, meaning, collocations and some examples as well as synonyms, where applicable. Information concerning other languages, the subject background, links etc. is not needed here, and therefore the dictionary will not offer such information.

The objective of lexicography is to describe, comprehensively or in specific segments, the vocabulary of a given language for specific purposes. (Klein & Geyken 2010:81)

It may be that the objective of linguistics is to describe a language. As regards dictionaries, this applies only to scientific documentation dictionaries. Other dictionaries do not aim to describe, but to assist the user who needs help with a communication problem that has occurred. For example, if someone is writing a Danish accounting annual report and is unsure whether *benchmark* is written with the Danish neuter article *et* or the common article *en*, he would look it up in the Danish Dictionary of Accounting's Text Production version (2011) and find the following:

**benchmark** <et, -et, -s, -ene>

This is not a description in the sense of a descriptive entry; it is a proscriptive entry (see Bergenholtz & Gouws (2010a) in this regard). In other words, it is not a prescriptive entry, but not a descriptive one either, as the common article is used. However, this variant is not recommended to the user, and he gets no further information from other entries in this dictionary either. The user receives the necessary reference which states that the variant with the neuter article is the variant recommended by this dictionary. In a further cognitive Danish Dictionary of Accounting: Knowledge (2011) the user would be able to access this information. He does not need this when he needs and looks for instruction on communication. It is neither necessary nor advisable to overburden the user with all the details in one and the same polyfunctional information tool:



The technically possible Utopia of universally accessible and complete documentation of all of mankind's knowledge can overburden individual users. (Haß & Schmitz 2010:4)

On the contrary: everything argues for the use of the technical possibilities solely to satisfy specifically certain kinds of needs for certain types of users.

It is indeed true that the current distinction between semantic and encyclopedia cannot be maintained, nor can the distinction between dictionary, lexicon and encyclopedia (cf. Bergenholtz & Kaufmann 1996). But the reason is not that it invites enthusiastic surfing. One can do that for the sake of lexicotainment, but for genuine information tools the rule is that they are better the more specifically and unambiguously they offer precisely the information the user wants in order to satisfy a particular information need, and nothing more than that.

On the whole, the terminology and the descriptive apparatus of linguistics are an unfamiliar as well as an irrelevant field to normal users of dictionaries. The user does not know these terms. If that is what the lexicographer (a linguist wearing a different hat) thinks, then the user indeed has to learn a lot before he can meet his information needs. The '*benchmark*' example given above does not prove the opposite. The common user is unlikely to know the term 'neuter article', he knows only the principle that there are two genders in Danish. Therefore, no provision is made for the user to press a button for declension morphology, for example, in order to obtain the information needed for the neuter gender. If he has such a problem, he should select the button "*I am writing a text*". The argument below therefore misses the user's needs and remains within the world of linguistics:

... consists of the series of components that can be edited and also used largely independently from one another ... a pronunciation module, a morphology module, a syntax module, a semantics module, an etymology module. (Klein & Geyken 2010:81)

Indeed, in all the dictionaries I have compiled so far there were different fields which the participating lexicographers had to fill with content. But the user was presented with a quite different classification, for example for a database with information for Danish: 1. I am reading a text, 2. I am writing a text knowing the word, 3. I am writing a text looking for the right word, 4. I want to know more about a word. This is language the users understand. They are unsure of the meaning of terms such as semantics, encyclopedia, etymology, inflection, syntax etc. For the same reason, most of the so-called user surveys are useless not only from a statistical perspective, since the leader of any survey that is to be taken seriously cannot be allowed to select the respondents himself, but should have them selected as a representative sample from a certain total population. They are also worthless especially because of questions of the type "Do you especially, or how often, look up in a dictionary in order to gather information on etymology,

syntax etc", as in Varantola (2002). As the respondents are usually students of linguistics, they understand these terms. However, such terms do not correspond to the basic needs of the person looking for information, such as "I am reading a text and do not understand the word".

#### 4. TWO DATABASES AS BASIS FOR MONOFUNCTIONAL DICTIONARIES

To formulate the present thesis more exactly: If one wishes to compile printed or electronic dictionaries these days, one needs first of all a database that should contain as many individual fields for specific data types as will be required in the dictionaries to be derived from it. Consequently, a database and a dictionary are not the same thing. They only appear to be almost the same thing if you turn it into a comprehensive polyfunctional information tool. As regards printed dictionaries, I regard this as optimal only in exceptional cases; as regards electronic dictionaries, I never do.

In the dictionary of music which was mentioned earlier, which has a very simple structure and only a few data fields, there are 13 fields which the programme searches and from which the relevant data for the respective dictionaries were also taken. Fields not displayed can be searched, but data from fields which were not part of the search can also be displayed. In this case exactly four dictionaries are abstracted from the database. The scheme for the reception dictionary is shown below. The search is conducted in only one of the fields, and it is done as a fuzzy search so that search results will be displayed even if the user has misspelt an expression. Data from only three fields are displayed in the order indicated. Lastly, provision is made for displaying a longer list of lemmas in case the search produces several results.

##### Meaning of Music Terms.

The button to click on is called "understand a musical term"

Search in fields	Field	Order of presentation	Listing
fuzzy search	lemma	1	1
	translation		
	example		
	language		
	abbreviation		
	short explanation	2	
	long explanation		
	remark		
	see also		
	synonym		

	Internet link		
	reference to the systematical introduction to music term		
	illustration	3	

In the next dictionary many more database fields are displayed. The database for the accounting dictionaries has a total of 84 fields; it contains fields related to accounting in the languages Spanish, Danish and English. The Danish-English Accounting Dictionary is a communication dictionary, and its function is to translate from Danish into English in the field of accounting. The button to click on is called "translation Danish-English". In this dictionary the programme does a minimising search, in other words the search does not proceed to the next step if the search in the first field searched has produced a result. First the exact lemma entered is searched for; then an inflected form of the lemma; then a lemma contained in the search string; and finally a fuzzy search follows if the first three search methods produced no result.

Here only 42 of the 84 fields of the databases are shown; all fields relating to Spanish have been omitted. The dictionary article which the user gets strongly resembles that of an ordinary translation dictionary. It should be noted, however, that there are for example synonyms for the English equivalents because, being a proscriptive dictionary (i.e. a recommending dictionary), only one equivalent entry is given for a meaning, namely the recommended one. Other possible equivalents are mentioned as further possibilities, but are not recommended as translations for the Danish lemma.

Search in fields	Field	Order of presentation	Listing
1. ==lemma; 2.==inflected form 3. *search string* 6. fuzzy search	1. Danish lemma	1	1
	2. jurisdiction label: Danish lemma	2	
	3. homonym index: Danish lemma	3	
	4. grammar: word class Danish lemma	4	
	5. grammar: inflectional paradigm Danish lemma		
	6. grammar note: Danish lemma		
	7. polysemy index: Danish lemma	5	

	8. definition of Danish lemma	6	
	9. synonym of Danish lemma		
	10. jurisdiction label: synonym of Danish lemma		
	11. antonym of Danish lemma		
	12. jurisdiction label: antonym of Danish lemma		
	13. "see also" related to Danish lemma		
	14. lexical note: Danish lemma		
	15. source of Danish lemma		
	16. links related to Danish lemma		
4. *search string*	17. collocation of Danish lemma + English collocational equivalent	11	
	18. jurisdiction label: Danish and English collocation	12	
5. *search string*	19. example of Danish lemma + English example	13	
	20. jurisdiction label: Danish and English example	14	
	21. "not recommended": Danish lemma		
	22. English equivalent	7	
	23. jurisdiction label: English equivalent		
	24. homonym index: English equivalent		
	25. grammar: word class English equivalent	8	
	26. grammar: inflectional paradigm English equivalent	9	
	27. grammar note: English equivalent		
	28. polysemy index: English equivalent		
	29. definition of English equivalent		
	30. synonym of English equivalent	10	
	31. jurisdiction label: synonym of English equivalent		
	32. antonym of English equivalent		
	33. jurisdiction label: antonym of English equivalent		
	34. "see also" related to English equivalent		
	35. lexical note: English equivalent		
	36. source of English equivalent		
	37. links related to English equivalent		
	38. English collocation		
	39. jurisdiction label: English collocation		

	40. English example		
	41. jurisdiction label: English example		
	42. "not recommended": English equivalent	15	

As the database contains many collocations for some accounting terms, the data input entry can become rather long in the case of important terms. For the translation of collocations, however, a different dictionary is more suitable in many respects – a dictionary that is also abstracted from the same large database. In total, there are 23 dictionaries; following the tradition of the printed dictionary, they are split up as follows: Accounting dictionaries for 1. Danish, 2. English, 3. Spanish, 4. Danish-English, 5. English-Danish, 6. English-Spanish and 7. Spanish-English. In each of these dictionary groups, buttons as explained above are provided which the user can then activate as needed.

## 5. CONCLUSION

There are two main tendencies in lexicography:

1. Lexicography as a linguistic discipline with a focus on linguistic methods and linguistic units. The theoretical consideration and the practical results presuppose the polyfunctional printed or electronic dictionary which intends to be consulted for many different purposes. But in the first place such dictionaries try to be a documentation tool for linguistic analyses.
2. Lexicography as a part of information science with a focus on certain user needs in certain types of user situations. Such information needs can at best be fulfilled in monofunctional information tools. This paper gives two examples of databases being the base for 4 resp. 23 different monofunctional dictionaries.

## REFERENCES

- Almind, R. & H. Bergenholtz. 2000. Die ästhetische Dimension der Lexikographie *Bild im Text – Text und Bild*, herausgegeben von U. Fix & H. Wellmann. Heidelberg: Winter. 259-288.
- Atkins, B.T.S. & M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Béjoint, H. 2010. *The lexicography of English*. Oxford: Oxford University Press.
- Bergenholtz, H. & M. Johnsen. 2005. Log Files as a Tool for Improving Internet Dictionaries. *Hermes. Journal of Language and Communications Studies* 34:117-141.

- Bergenholtz, H. & M. Johnsen. 2007. Log files can and should be prepared for a functionalistic approach. *Lexikos* 17:1-20.
- Bergenholtz, H. & R. Gouws. 2007. Korrek, volledig, relevant. Dít is die vraag aan leksikografiese definisies. *Tydskrif vir Geesteswetenskappe* 47(4):568-586.
- Bergenholtz, H. & R. Gouws. 2010a. A Functional Approach for the Choice between Descriptive, Prescriptive and Proscriptive Lexicography. *Lexikos* 20:26-51.
- Bergenholtz, H. & R. Gouws. 2010b. A new perspective on the access process. *Hermes. Journal of Language and Communications Studies* 44:103-127.
- Bergenholtz, H. & S. Tarp. 2002. Die moderne lexikographische Funktionslehre. Diskussionsbeitrag zu neuen und alten Paradigmen, die Wörterbücher als Gebrauchsgegenstände verstehen. *Lexicographica* 18:253-263.
- Bergenholtz, H. & S. Tarp. 2003. Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *Hermes. Journal of Language and Communications Studies* 31:171-196.
- Bergenholtz, H. & U. Kaufmann. 1996. Enzyklopädische Informationen in Wörterbüchern. *Semantik, Lexikographie und Computeranwendungen*, herausgegeben von N. Weber. Tübingen: Niemeyer. 168-182.
- Bergenholtz, H. 1996. Grundfragen der Fachlexikographie. *Euralex '96. Proceedings I-II. Papers submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden*, edited by M. Gellerstam & J. Järborg & S.-G. Malmgren & K. Norén & L. Rogström & C. Røjder Papmehl. Göteborg: Göteborg University. 731-758.
- Bergenholtz, H. 1998. Das Schlaue Buch. Vermittlung von Informationen für textbezogene und textunabhängige Fragestellungen. *Symposium on Lexicography VIII. Proceedings of the Eighth International Symposium on Lexicography May 2-5, 1996 at the University of Copenhagen*; edited by A. Zettersten & J.E. Mogensen & V. Hjørnager Pedersen. Tübingen: Niemeyer. 93-110.
- Bergenholtz, H. 2009. Schnellerer und sicherer Datenzugriff in gedruckten und elektronischen Fachwörterbüchern und Lexika. *Revue française de linguistique appliquée, dossier: terminologie orientations actuelle* 14(2):81-97.
- Danish Dictionary of Accounting: Knowledge. 2011. = S. Nielsen & L. Mourier & H. Bergenholtz. *Den Danske Regnskabsordbog: Viden*. Database: Richard Almind. Odense: Ordbogen.com. ([www.ordbogen.com](http://www.ordbogen.com)).
- Danish Dictionary of Accounting: Text Production. 2011. = S. Nielsen & L. Mourier & H. Bergenholtz. *Den Danske Regnskabsordbog: Tekstproduktion*. Database: Richard Almind. Odense: Ordbogen.com. ([www.ordbogen.com](http://www.ordbogen.com)).

- Danish-English Accounting Dictionary. 2011. = S. Nielsen & L. Mourier & H. Bergenholtz. *Den Dansk-Engelske Regnskabsordbog*. Database: Richard Almind. Odense: Ordbogen.com 2011. (www.ordbogen.com)
- Fuertes-Olivera, P., S. Nielsen, H.J. Bergenholtz, L. Mourier, P. Gordo Gómez, M.N. Amo, A. de los Rios Rodicio, A. Sastre Ruana, S. Tarp, M.S. Velasco Sacristán & R. Almind. 2010. *Diccionario de Contabilidad Inglés-Español*. Madrid: Thomson Reuters-Aranzadi.
- Klein, W. & A. Geyken. 2010. Das Digitale Wörterbuch der Deutschen Sprache (DWDS). *Lexikographica* 26:79-96.
- Leth, J. 1800. *Dansk Glossarium. En Ordbog til Forklaring over det danske Sprogs gamle, nye og fremme Ord og Talemaader for unge Mennesker og for Ustuderede. Et Forsøg. Med en Fortale af Professor Rasmus Nyerup*. København: Trykt paa Hofboghandler Simon Poulsens Forlag hos Bogtrykker Morthorst's Enke & Comp.
- Meaning of Music Terms. 2011. = I. Bergenholtz in cooperation with R. Almind and H. Bergenholtz. *Betydning af musikudtryk*. Odense: Ordbogen.com. (www.ordbogen.com).
- Nielsen, S. & L. Mourier & H. Bergenholtz. 2007. *Regnskabsordbogen engelsk-dansk*. Copenhagen: Thomson.
- Roux, J.C. 2011. Electronic dictionaries for speech recognition, for its applications and for speech synthesis. *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*, edited by R.H. Gouws & U. Heid & W. Schweickard & H.E. Wiegand. Berlin: Mouton de Gruyter (in print).
- Varantola, K. 2002. Use and usability of dictionaries: common sense and context sensibility. *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkin*. edited by M.-H. Corréard. Grenoble, France: EURALEX.30-44.

## CHAPTER 20

### **'MARKETSPEAK' IN IGBO: A SPEECH SYNTHESIS TRAINING PROJECT**

Dafydd Gibbon<sup>1</sup>, Ugonna Duruibe<sup>2</sup>, Jolanta Bachan<sup>3</sup>

*<sup>1</sup>Bielefeld University, Bielefeld, Germany*

*gibbon@uni-bielefeld.de*

*<sup>2</sup>University of Ibadan, Ibadan, Nigeria*

*geenah22@yahoo.com*

*<sup>3</sup>Adam Mickiewicz University, Poznań, Poland,*

*jolabachan@gmail.com*

#### **1. OBJECTIVE: TECHNOLOGY TRAINING AND THE 'MARKETSPEAK' SCENARIO**

The objective of the present contribution is to provide a missing tutorial link in current discussions of speech technology for less resourced languages with little high quality data and incomplete descriptions, on the one hand, but with local 'human resources' who are not specialists in speech technology, but are trained either in linguistics or in computer science. The scientific primacy of theoretical and descriptive novelty – the 'syntax' and 'semantics' of science – is of course uncontested, but here we deal with the 'pragmatics' of science and technology applied to education in the field of speech technology in an environment with restricted infrastructure.

The specific task pursued in the present context is the development of components of a speech synthesiser for use with the Nigerian language Igbo (Benue-Congo, ISO 639-3 *ibo*) in the context of marketing goods and prices, the 'MarketSpeak' scenario, as a model for generic solutions in this field. Specifically, the present contribution presents an outcome and further development of a workshop on speech synthesis for Nigerian languages, in Abuja, Nigeria, March 2010, sponsored by a major project<sup>1</sup> on generic text-to-speech applications for African Tone languages.

The workshop goals were to provide basic training in speech synthesis for a mixed group of linguists and computer scientists who were not specialists in speech technology. Speech synthesis was chosen as a more feasible entry into speech technology than speech recognition based technologies. The specific goal of this tutorial was to create prototype 'microvoices'<sup>2</sup> for speech synthesis of 12 Nigerian languages by the participants, who were native speakers of these languages; this goal was achieved.



After the workshop, a synthetic voice front end for Igbo was created (Duruibe 2010), specifically for the restricted register of Igbo food markets, using straightforward traditional technology for diphone synthesis<sup>3</sup>. The aim was to create a first voice on which to build a rule-based Text-to-Speech (TTS) synthesiser for Igbo.<sup>4</sup>

The present study reports on this work, as we feel that it offers a novel approach to teaching the basics of viable speech synthesis to non-specialists. For this purpose we introduced new heuristic procedures for automatically creating a phonetically rich data set for recording, for automatically extracting diphones from speech data, and for evaluating data quality and system quality by providing close copy gold standard benchmarks. The new contribution of the present study lies not in the development of novel theories, models, algorithms and application domains (Roux *et al.* 2010) but in novel combination and deployment of known technologies in new fields of application for less resourced languages, for non-specialists with basic linguistic and/or computing knowledge. In this context, teaching strategies and low budget speech synthesis development methods must be used, with the focus on a community with a language which so far has few empirical, descriptive and technological resources but a strongly felt need for and interest in technological development.

An optimal solution for the task, if it were solely system development and not an educational task, would be an easy-to-use speech synthesis kit with clear linguistic interfaces and user-friendly tools for data selection, processing and evaluation, but unfortunately, to date there is no such kit. Consequently, criteria for using results from different areas of linguistics, computational linguistics and speech technology were integrated for this purpose. The present contribution concentrates on the requirements and creation of the phonetic and digital signal processing (DSP) components of a text-to-speech synthesiser. The Natural Language Processing (NLP) components and their computational linguistic foundations are only dealt with in passing.

In Section 2, software requirements and design are discussed, followed by discussion of linguistic specifications of the system prototype in Section 3. In Section 4 the pre-recording, recording and post-recording phases of data processing for voice development is dealt with in some detail, and the workflow is presented. Section 5 briefly describes the voice construction step, and Section 6 presents conclusions and an outline of future work.

## **2. REQUIREMENTS**

### **2.1. Software Selection Requirements**

The general requirements for providing a feasible method for rapid speech synthesiser development for Nigerian tone languages have already been outlined. The general requirements lead to a number of specific requirements which determined the choice of the speech synthesis method. The choice was a conservative one, which fell on the MBROLA diphone synthesiser, for the following reasons:

1. Suitability for use in a training context with minimal training time for linguists whose knowledge of computation is limited to using consumer software, and computer scientists with no more than a basic knowledge of phonetics.
2. Free software, because of minimal or no funding.
3. Comprehensive documentation (Dutoit, Pagel 1996; Dutoit 1997<sup>5</sup>), to facilitate understanding of procedures and to encourage further creative development.
4. Credentials of extensive use for multilingual speech synthesis (73 voices for 36 languages are publicly available on the internet).
5. Cross-platform availability (runtime binaries for 37 operating systems and operating system versions are available, including the required Linux and Windows versions).
6. Offline use, independence from internet tools, because of the expense and unpredictability of internet access.
7. Ease of installation, to facilitate deployment by non-specialists.
8. Simple interface between text parser and diphone synthesis components, to permit close cooperation between linguists, computational linguists and phoneticians.
9. Reasonable quality in relation to the other requirements.

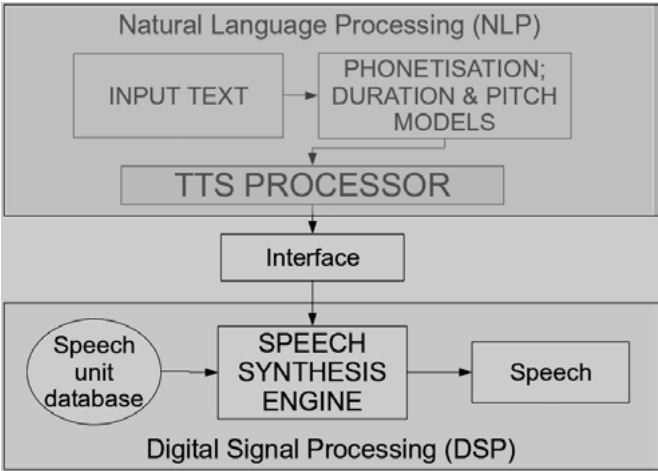
Clearly MBROLA is not a state of the art system any longer, but there is no other speech synthesis system which fulfils requirements 1-8 above to anything approaching the extent to which they are fulfilled by MBROLA, though there are better quality synthesisers. Although 36 languages are represented in the public MBROLA voice collection, the only African language represented there is Afrikaans, which is historically and typologically unique on the continent in being

closely related to European languages. Thus there is a lot of room for further development and experimentation in relation to typologically different African languages.

2.2. System Architecture Requirements

Figure 1 shows the overall architecture within which the speech synthesis front-end for Igbo is designed. Currently the focus is on the diphone synthesiser DSP front-end, i.e. the component which produces acoustic output from a linguistically oriented representation of pronunciation, containing both segmental and prosodic information. The computational linguistic foundations of the back-end NLP components of preprocessing, parsing and the automatic creation of pronunciation models were not created in the prototype, and for this reason the NLP back-end is greyed out in Figure 1, though some aspects are briefly outlined in Section 3, in the discussion of linguistic specifications for Finite State grammar and tone modelling.

Figure 1: General architecture model for text-to-speech synthesis.



The important feature of the MBROLA concept, which is not always found in other speech synthesis concepts, is the *Interface* specification, which permits both segmental units and their prosodic properties to be specified in a linguistically transparent fashion. The interface is implemented as a text file with the extension “.pho” (informally referred to as a ‘pho file’). The structure of the pho file representation is defined as follows:

```

<pho-file>      ::= <pho-line>*
<pho-line>      ::= <commentline> | <phonemespec> | emptyline
<phonemespec>   ::= phoneme duration <pitchspec>*
<pitchspec>     ::= position hertz
<commentline>   ::= # char*

```

In other words, the pho file representation consists of an arbitrary number of lines (actually depending on the length of the utterance to be synthesised), which may either be a comment starting with a hash character "#", a phoneme specification, or an empty line.

The phoneme specification contains three kinds of information:

1. A *phoneme label*, usually in the keyboard-friendly SAMPA encoding of the International Phonetic Alphabet (cf. Gibbon *et al.* 2000a:359ff., also available via internet search). Sometimes major allophones are specified in addition to phonemes.
2. A *duration value in milliseconds*, specifying the length of the phoneme in the context of its preceding and following neighbours. The duration value is in general determined by statistical analysis of durations in a corpus of annotated speech, often by means of a classification and regression tree (CART) which weights different factors found in the corpus.
3. A *pitch contour specification*, consisting of a series of zero or more *pitch value specifications*. Each pitch value specification consists of a *position specification* as a percentage of the phoneme/phone (early, e.g. 20%, mid, i.e. 50%, late, e.g. 80%) and the *pitch specification* in Hertz. By supplying a sequence of specifications, tonal contours can be emulated. Voiceless sounds and pauses are generally not supplied with a pitch value specification.

The interface content specification determines not only the data requirements for the recording phase of system development but also the requirements for phonetic information:

1. The phoneme inventory of the language concerned (perhaps with major allophones).
2. A duration model for phoneme lengths.
3. A pitch model for specifying the shapes of contours on specific phonemes.

An example of an interface specification for Igbo is shown in Table 1.

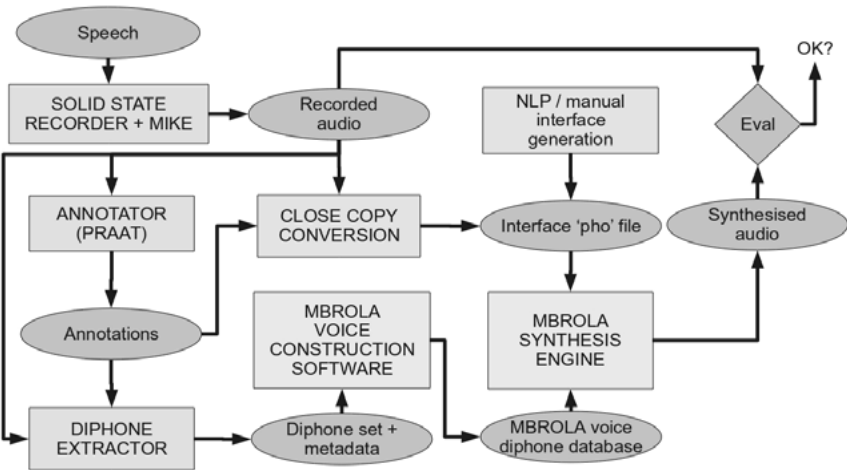
**Table 1:** *Input interface table to MBROLA diphone synthesiser front-end.*

Phoneme (SAMPA)	Duration (ms)	Pitch specification				...	...
		%pos	Hz	%pos	Hz		
–	200						
a	301	80	180				
gw	223						
a	169	80	145	(only one pitch specification per line in this file)			
–	1445						
O	331	80	234				
k	162						
a	231	80	145				
–	1296						

**2.3. System Workflow Requirements**

The system architecture is centred on the interface between the NLP and the DSP components, and on the construction of phoneme models with their duration and pitch characteristics. The phoneme model is taken from linguistic analyses of Igbo (see below), and the duration and pitch characteristics are taken from computational corpus analysis of the recordings, implemented with Python and Unix/Linux shell scripts. The overall workflow which is required for producing the Igbo voice is derived from the architecture, and is shown in Figure 2. The figure is intended to be self-explanatory, given the previous discussion. However, for clarification the main inputs and workflow phases are described more fully in context.

**Figure 2:** *Workflow for Igbo diphone voice development.*



First, speech input for voice development is digitally recorded, and used for 3 purposes:

1. Evaluation – the ultimate ‘gold standard’ for comparison with the voice output.
2. As the basis for the (manual) creation of annotation files, e.g. with the Praat toolbox (Boersma 2001; Boersma & Weenink 2011):
  1. for use in the automatic (or manual) creation of close-copy synthesis input with the voice, for ‘gold standard’ evaluation. For this purpose, the phonemes and their durations are extracted from the annotation, and the pitch values are extracted from the speech signal, and formatted in pho file interface format;
  2. for use, together with the speech signal, in automatically extracting diphones from the speech signal by means of a Python script, as input for the MBROLA voice creation procedure.

Second, the diphone set is created from the corpus. Using Python scripts, the metadata about the diphone timestamps in the speech database are extracted from the speech annotations, and diphone files are extracted from the speech file for input to the MBROLA voice construction software (the ‘Mbrolator’), which processes the diphone files into an MBROLA voice.

Third, the inputs to the MBROLA runtime synthesis engine are a pho file interface (created manually, or automatically with an NLP component), and the MBROLA voice. The output of the synthesiser is evaluated in perception tests, in which it is compared with utterances from the original speech database.

Finally, the voice is evaluated in perceptual tests using the pho file representations based on close copies of the original speech and on automatically generated versions.

### **3. LINGUISTIC SPECIFICATIONS**

#### **3.1. Grammar Component: Nominal and Verbal Sequences**

For the purpose of modelling grammar in the present scenario a Regular (Type III) Grammar or Finite State Automaton (FSA) model is assumed to be adequate. The conditions for which a general context-free model is needed, i.e. centre-recursive sentence embedding of expressions, are more typical of formal written text, and for practical purposes can be excluded from models of restricted spontaneous speech.

Examples of finite state models for typical nominal and verbal sequences found in Benue-Congo languages like Igbo are shown in Figure 3 (Gibbon *et al.* 2003). The models were originally developed for the neighbouring and both historically and typologically closely related Benue-Congo language Ibibio (ISO 639-3: *ibb*).

**Figure 3:** *Ibibio Noun Phrase (left) and Verb (right) sequences.*

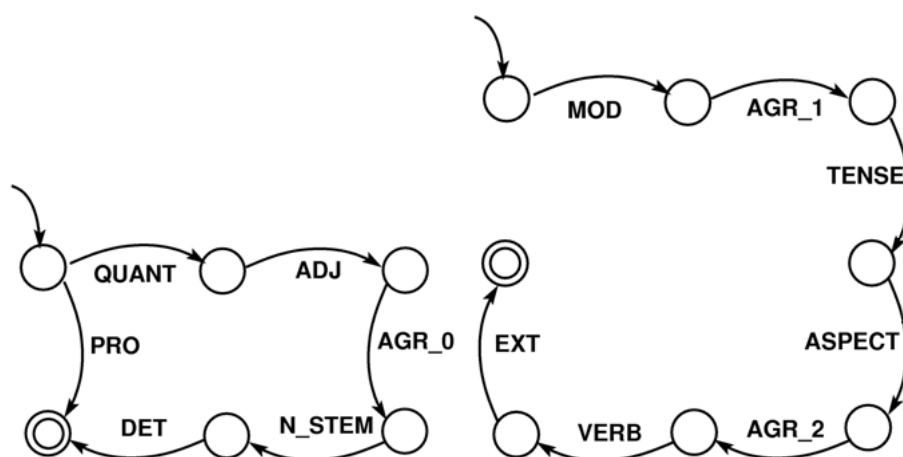


Figure 3 shows two transition diagrammes representing one Finite State Automaton (FSA) for Noun Phrases and one for Verbs. Transition diagrammes can be seen as a kind of map, in which the states (represented by circles) symbolise places and the transitions (represented by arrows) represent events involving motion from one place to another (e.g. recognition or generation of a word). The starting point is represented by an arrow with no circle attached to the beginning, and terminal states are represented by a double circle. Classes of words, rather than individual words, are shown, labelled with grammatical categories.

The Noun Phrase transition diagramme shows that a Noun Phrase can be either a PROnoun, or a sequence of QUANTifier, ADJective (optionality not shown), a Subject AGReement morpheme (AGR\_0), a Noun stem and a DETerminer morpheme. In addition to lexical tone, each of these elements is modified by a tonal specification.

The Verbs are agglutinative: a stem concatenated with sequences of prefixes and suffixes. The Verb transition diagramme shows that Verbs consist of a MODality morpheme, an AGReement morpheme (AGR\_1) which agrees with the Subject AGReement morpheme (AGR\_0), a TENSE morpheme, an ASPECT morpheme, an Object AGReement morpheme which agrees with the AGR\_0 morpheme of the Object Noun Phrase, and an EXTension with various functions. In addition to

lexical tone, each of these elements is modified by a grammatical tone specification, and tone agreement is subject to further constraints.

Initial investigation shows that Igbo and Ibibio grammar do not differ greatly in these respects, though the vocabulary and detailed constraints certainly differ. When an NLP component is added to the present voice prototype, it will be based on a grammar of this type.

### 3.2. Tone Sequences

The tone sequencing properties of Igbo involve tone terracing, a special case of the downtrends outlined by Connell (2002). The fundamental frequency (F0) values associated with each individual phoneme are not lexical properties of the phoneme but dependent on several factors, expressed by a function with at least the following six factors:

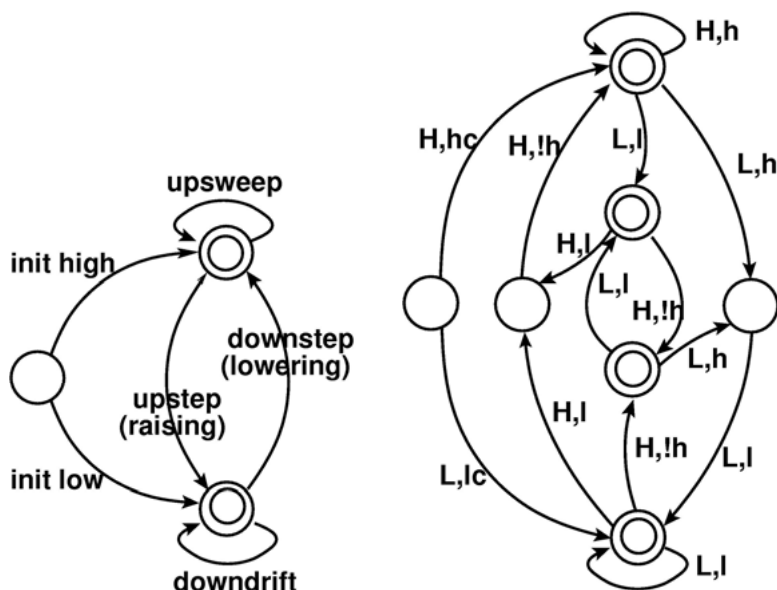
$$F0(\text{phoneme}_i) = f(\text{baseline}, \text{onsetpitch}, \text{declination}_i, \text{perturbation}, \text{tone}, \text{intonation})$$

The index  $i$  indicates the position of the phoneme in the sequence. The *baseline* factor is a value below which the frequency does not fall. The *onsetpitch* factor is the initial pitch minus the baseline. The *declination<sub>i</sub>* factor is generally <1 and the power superscript determines an asymptotically downtrending frequency trajectory, depending on intonation factors, e.g. in questions this factor can also be >1, or have an additive pitch-raising element. The *perturbation* factor is the effect of the modification or blocking effect of the consonantal or vocalic segment *phoneme<sub>i</sub>* on the frequency. The *tone* factor is the lexical or grammatically determined contrastive tone (cf. Gibbon *et al.* 2009; this factor corresponds structurally to *accent* in a stress/accent language). The *intonation* factor is a complex global pattern associated with speech acts and focus (Hirst *et al.* 1998; for Igbo cf. Ikekeonwu 1993). The *baseline*, *onsetpitch* and *declination<sub>i</sub>* factors together determine the overall terracing downtrend. For related work, cf. Liberman *et al.* (1993), Pierrehumbert and Liberman (1994), Akinlabi and Liberman (2000).

The finite state model which accounts for the basic tone terracing pattern for Niger-Congo languages is shown in Figure 4. The simple model applies in principle to Igbo, but does not apply equally to all Niger-Congo languages (even apart from toneless cases such as Swahili): the case of Baule (Kwa, ISO 639-3 bci) is an example of a case in which added complexity is required in order to account for three-tone sequence effects rather than effects between two neighbouring tones (Gibbon 1987; cf. also Gibbon 2001, 2009 for further details).



**Figure 4:** *Tone sequence models: General tonological model for Niger-Congo 2-tone languages and specific model (with lookahead) for Baule*



In the phonetic interpretation of the categories represented in the transition diagrammes of Figure 4, patterns such as downstep, downdrift, upstep and upsweep are modelled by operations on the pitch of the immediately preceding Tone Bearing Unit (TBU), relative to a reference pitch baseline. The modifications are logarithmic, implemented by multiplication of the preceding pitch value by a factor  $< 1$  or  $> 1$ , depending on High or Low tone, with a cumulative logarithmic effect in the course of the utterance. In principle, the procedure resembles that described for Igbo by Liberman *et al.* (1993) and formulated by Akinlabi and Liberman (2000) as follows:

1. Divide the utterance into maximal regions of like tone [represented in the automata by the local loops - Authors].
2. Place a mid-valued tonal target at the start of the utterance [the 'init\_low' and 'init\_high' events - Authors].
3. Place a tonal target at the end of each region, choosing an F0 value determined by the tonal type, downdrift/downdstep, and final boundary effects if any [not necessary in the automata - Authors].
4. Interpolate linearly from target to target [the events on the loop transitions - Authors].

However, the particular non-local 'lookahead' strategy selected by Akinlabi and Liberman is not confirmed in phonetic investigations on sequences of like tones in neighbouring Ibibio (Gibbon *et al.* 2000b), where a logarithmic approach is more appropriate, and better represented by the local 'immediate neighbour' strategy of the automata and their phonetic interpretation. It is an empirical question whether Igbo differs so much from Ibibio in this respect.

## 4. DATA

### 4.1. Data and Data Processing Requirements Specification

The data and data processing requirements are derived from the task and architecture specifications and may be summarised as follows:

1. Pre-recording phase: the scenario vocabulary, the scenario prompts for recording, definition of the phoneme inventory, definition of the diphone inventory.
2. Post-recording phase: annotation of recordings; extraction of diphone time-stamps; extraction of diphone files; assignment of phonemes, durations and pitch specifications in pho files.

### 4.2. The Scenario Prompts (Wordlist and Sentence List)

The prompt list comprises words from the semantic fields *Food Items*, *Domestic Animals*, *Fruits and Vegetables*, *Kitchen Utensils*, *Other Items*, and *Market Days*, as well as assorted sentences, including formulaic greetings. To illustrate the lexical data, the word field *Domestic Animals* is shown in Table 2.

**Table 2:** Sample Igbo vocabulary database table from semantic field 'Domestic Animals'.

Orthography	IPA transcription	SAMPA transcription	Corpus Tones	Gloss
ehi	/ehi/	ehi	H-H	'cow'
ewu	/ewu/	ewu	H-H	'goat'
atūrū	/aturū/	atUrU	H-H-!H	'lamb'
ọkūkọ	/ɔkukɔ/	OkUKO	L-H-L	'fowl'
ezi	/ezi/	ezi	H-L	'pig'
anụ	/anu/	anU	H-H	'meat'
ejule	/edʒule/	edZule	H-L-L	'snail'

4.3. Phoneme Inventory

The phoneme inventory of Igbo is known (cf. Duruike 2010) and does not need to be specified phonetically here. More important is the presence of an adequate number of phoneme instances. A phoneme frequency list for the ‘MarketSpeak’ corpus is shown in Table 3. Including pause, Igbo has 38 phonemes according to the analysis used here for Igbo voice construction: /a/, /b/, /tʃ/, /d/, /e/, /ɛ/, /f/, /g/, /ɸ/, /ɣ/, /gʷ/, /h/, /i/, /ɪ/, /dʒ/, /k/, /kʷ/, /l/, /m/, /n/, /ŋ /, /ɲ/, /ŋʷ/, /o/, /ɔ/, /p/, /pʰ/, /r/, /s/, /ʃ/, /t/, /u/, /ʊ/, /v/, /w/, /j/, /z/, plus pause. The phonemes are represented for computational purposes with the X-SAMPA IPA encoding for keyboarding convenience (cf. Gibbon *et al.* 2000a:359ff.); the pause is represented by the understroke “\_”.

For the present contribution, justification of the details of the descriptive phonetic definitions of specific phoneme labels is less significant than having a complete inventory of phonemes together with accurate labelling of the recorded speech signal, because the phonemes are mapped in diphone contexts directly to the acoustic representation obtained from the corpus, and not to a phonetic representation of the traditional symbolic type, thus capturing local allophonic variation.

Interestingly, this direct mapping from phonemes to acoustic representations corresponds exactly to the view of Bloomfield (1933), who did not consider the intermediate level of phonetic representation in symbols to be particularly important for the development of a realistic theory.

**Table 3:** *Corpus frequency list for ‘MarketSpeak’ Igbo microvoice.*

112	_	24	n	11	z	5	s	3	j
76	a	24	E	10	kw	4	p	2	NX
39	U	23	o	9	tS	4	Nw	1	w
37	O	21	k	9	dZ	4	J	1	v
37	l	19	r	7	h	4	gw	1	S
29	e	17	u	7	b	4	f	1	G
28	i	14	d	6	l	4	bY		
25	m	13	g	5	t	3	pY		

The upper bound for the number of diphones required for a given language is given by the square of the size of the phoneme inventory (including the ‘pause phoneme’). There are 38 phonemes in the corpus, including pause, so the upper bound for the number of intra-word and inter-word diphones required for a full corpus is 38 squared, i.e. 1444. A total of 253 diphones are represented in the

'MarketSpeak' corpus, i.e. about 20% of the upper bound. The diphones, sorted by frequency, are shown in Table 4. Even if the constraints on Igbo phoneme co-occurrence are such that the full complement of 1444 diphones is not attested, the number of unique phoneme occurrences shows that the entire potential of the Igbo diphone set is not reached. It is therefore clear from the frequency tables for the corpus that, while sufficient for microvoice testing purposes in the selected scenario, the current diphone coverage of this corpus will not provide a complete Igbo voice.

**Table 4:** Igbo diphone frequency list from 'MarketSpeak' corpus.

Freq	Diphone	Freq	Diphone	Freq	Diphone	Freq	Diphone	Freq	Diphone
28	a _	3	o r	2	i Nw	1	pY U	1	h E
19	_ a	3	O m	2	l n	1	pY O	1	h a
16	U _	3	o l	2	l m	1	pY o	1	gw u
12	O _	3	O k	2	l k	1	p i	1	gw O
12	_ o	3	n E	2	i e	1	O z	1	gw e
12	e _	3	m U	2	l d	1	o v	1	gw a
11	_ O	3	l e	2	h l	1	O tS	1	G l
11	l _	3	J a	2	h i	1	O pY	1	g e
10	_ U	3	l tS	2	g l	1	o pY	1	e z
10	i _	3	g U	2	g i	1	O p	1	e w
10	E _	3	g o	2	g a	1	o p	1	e r
10	_ e	3	E r	2	_ g	1	o o	1	e O
8	m a	3	dZ l	2	f O	1	O l	1	E Nw
8	_ m	3	d U	2	f E	1	o k	1	E kw
8	k a	3	b l	2	e n	1	O h	1	E k
8	a n	3	a J	2	E m	1	O f	1	e k
7	m m	2	U z	2	e h	1	n z	1	e i
7	_ i	2	U t	2	e g	1	NX U	1	e gw
6	z U	2	U r	2	E f	1	NX a	1	E G
6	_ u	2	U n	2	e dZ	1	_ Nw	1	E g
6	u _	2	u n	2	E d	1	n tS	1	e d
6	_ n	2	U bY	2	d e	1	n r	1	e a
6	_ l	2	tS a	2	a r	1	n l	1	dZ u
6	_ E	2	t e	2	a p	1	n gw	1	d u
5	o _	2	s e	2	a NX	1	m e	1	d o
5	kw a	2	r U	2	a kw	1	m b	1	d i
5	k l	2	r a	2	a l	1	m _	1	d E
5	l a	2	o m	2	a h	1	l u	1	d a
5	dZ i	2	O kw	1	z u	1	kw U	1	bY u
5	a z	2	O g	1	z O	1	kw E	1	bY O
4	U O	2	o g	1	z l	1	kw e	1	bY e
4	r O	2	O dZ	1	z i	1	k U	1	bY a
4	O r	2	O d	1	z a	1	k O	1	b U

Freq	Diphone	Freq	Diphone	Freq	Diphone	Freq	Diphone	Freq	Diphone
4	n U	2	Nw E	1	w u	1	J E	1	b O
4	i k	2	Nw a	1	v a	1	j a	1	b E
4	l b	2	n u	1	u s	1	_ J	1	b e
4	_ dZ	2	n n	1	u r	1	l z	1	a t
4	d l	2	n kw	1	u l	1	l S	1	a pY
4	a tS	2	n g	1	U kw	1	i s	1	a O
4	a k	2	n d	1	U l	1	l r	1	a m
3	U k	2	n a	1	u h	1	i r	1	a l
3	u d	2	m j	1	U g	1	l O	1	a j
3	t U	2	m l	1	U d	1	l kw	1	a gw
3	tS O	2	l a	1	u bY	1	i kw	1	a g
3	tS l	2	kw O	1	U b	1	i i	1	a f
3	s i	2	k o	1	u b	1	l h	1	a E
3	r o	2	k E	1	U a	1	i gw	1	a dZ
3	r i	2	k e	1	tS i	1	i d	1	a a
3	r E	2	_ k	1	S a	1	i bY	1	- -
3	p a	2	j O	1	r u	1	i a		
3	o s	2	i t	1	r l	1	h u		

In order to create a complete voice, a number of strategies are available for creating the required data set:

1. *Wordlist Strategy*: Create a list of words and word combinations which exhaust the number of possible combinations of 2 phonemes, and place these in a context frame corresponding to the traditional “Say \_\_\_\_ again.” This traditional method has many disadvantages, for instance creating artefacts resulting from boredom and ordering effects.
2. *Natural Context Strategy*: Create a set of sentences containing at least one token of all possible combinations. This method has the advantage of more realistic and interesting prompts and is less likely to contain artefacts of the kind noted under the Wordlist Strategy.
3. *Minimal Natural Context Strategy*: From the set of sentences defined for the second strategy, select the smallest set containing all the diphones (Bachan 2010). It may be possible to reduce the number of prompts by this method to only a few hundred. Given a large corpus of sentences, this can be done with a software tool; final checking for missing diphones is always necessary. This method has the advantage of producing a smaller corpus than the Natural Context Strategy for annotation, a time-consuming and expensive activity, even if automatic segmentation with post-editing is available.

The main approach taken here is the Wordlist Strategy, due to the small size of the scenario-determined vocabulary, supplemented with the Minimal Natural Context Strategy for multi-word expressions.

#### **4.4. Duration Specifications**

The calculation of appropriate durations for sounds in different contexts is a complex issue, and was dealt in a highly simplified fashion in the present work by simply automatically averaging the durations for each phoneme as specified in the annotation time stamps, or, in the case of close copy tests, by automatically copying the durations from the annotation time stamps. A kind of null case was also generated, using arbitrary uniform segment lengths of around 80ms, which surprisingly led to very comprehensible and reasonably natural utterances. On reflection, this initially surprising result could, however, be a function of what is apparently a fairly uniform syllable timing in Igbo, another empirical question for future work.

#### **4.5. Pitch Specifications**

For close-copy tests there was no problem in automatically copying the sampled F0 from the original recordings, averaging the samples, and using these in the PHO representations. The results were pretty much identical to the original recordings.

A null case was also generated, based on the frequency model introduced above, retaining the asymptotically descending terracing function:

$$F0(\text{phoneme}) = f(\text{baseline}, \text{onsetpitch}, \text{declination})$$

This null case is 'tone-deaf', i.e. with no modification by lexical or grammatical High, Downstepped High and Low tone. Consequently, the contour was predictably not very natural, but still comprehensible, relying on native speaker ability to disambiguate in the case of tonal minimal pairs for which the tonal cues were missing. The practical value of this null contour in the present tutorial context was initially to provide a starting point for manually adding local modifications of High tone, Downstepped tone, Low tone and utterance-final lowering, for training and evaluation purposes.

The automatic assignment of lexical and grammatical and intonationally determined pitch values is a front end issue and outside the scope of the present account; for test purposes, the values were calculated straightforwardly with a Python script.

## 5. DIPHONE DATABASE CONSTRUCTION

Given the recordings and the annotations, a number of software tools were used to construct the diphone database (voice).

In order to extract the diphones from the recordings, a suite of Python scripts developed by the third author was used: the first creates a table of diphones and calculates their time-stamps from information in the annotations; the second creates the diphone file set and the table of diphone information required by the Mbrolator. The format is shown in Table 5.

**Table 5:** *Input format for Mbrolator software tool set.*

Diphone filename	Diphone		Start	End	Mid
a-ee_Ugolgbo01_599.wav	a	E	800	2549	1588
a-ii_Ugolgbo01_537.wav	a	I	800	2332	1526
a-jj_Ugolgbo01_336.wav	a	J	800	3073	1813
a-nnX_Ugolgbo01_269.wav	a	NX	800	2800	1882
a-oo_Ugolgbo01_592.wav	a	O	800	3030	1856
a-SIL_Ugolgbo01_4.wav	a	_	800	3751	2151
a-a_Ugolgbo01_526.wav	a	A	800	2365	1510
a-dzz_Ugolgbo01_322.wav	a	dZ	800	3358	2222
a-f_Ugolgbo01_469.wav	a	F	800	4486	2527
a-g_Ugolgbo01_182.wav	a	G	800	2652	1696

The Mbrolator tools were provided under licence by the developers, and consist of a library of signal processing routines, and three user accessible tools: one for setting basic conversion parameters, one for processing the parameters, and one for creating the database on the basis of these parameters. The input to the Mbrolator tools consists of the set of diphones in separate files, an information file containing phoneme specifications, and a table that also contains a list of diphones with starts, mid and end time-stamps from the diphone files. The output is a database in the form of a single file containing the normalised diphones and information from the information file.

Full details for use of the software tools are given on the MBROLA website and in Bachan (2007 & 2010), Bachan *et al.* (2006) and Gibbon *et al.* (2008).

## 6. CONCLUSION AND PROSPECTS

A toolset and a workflow designed for speech technology education purposes were created with the aim of filling a gap in the available applied linguistic and speech technology literature, with the practical goal of simplifying the creation of basic synthetic voices for restricted scenario speech synthesis applications for under-resourced languages. On this basis, a working prototype DSP component for Igbo speech synthesis was created and satisfactorily evaluated using standard methods (Gibbon *et al.* 1997).

The choice of a traditional diphone synthesiser was motivated in detail, and the essential stages of the workflow were presented. For the Igbo 'MarketSpeak' scenario a set of prompts was created, recorded and annotated, and diphones and metadata about the diphones were extracted automatically from the annotated recordings, and a voice was created.

As noted during the discussion, the NLP component and its computational linguistic foundations were only treated in passing. There are three main descriptive linguistic and computational linguistic complexities which are not accounted for by the models discussed here: grammatical and lexical tone mapping, segment-tone interaction, and influence of sentence-level intonational categories such as focus (which can disrupt local downstep) and questions (which can disrupt overall declination); cf. Ikekeonwu (1993). There are models available which appear to be suitable for these purposes, such as multi-tape finite state transducers, which have been used for related languages (Gibbon *et al.* 2006b), and were sketched in Figure 3. These issues remain for future work.

Many attempts have been made to provide natural language and spoken language resource and toolkit specifications for under-resourced languages, as in the BLARK, the Basic LAnguage Resource Kit (Krauwier 2005), and extensions such as those of Gibbon *et al.* (2006a). However, these resource kits have not been specifically designed for use in training and basic development situations by linguists and computer scientists with no intensive speech engineering training, but with the plain motivation to create practical speech synthesis applications for their languages.

There have also been many specific state-of-the-art speech technology applications created for previously under-resourced languages in many countries across the world, including African languages. These cutting edge applications represent great strides forward in the field, and in specific application areas, but so far are only usable by specialists. There have been initiatives for combining cutting edge methodology with ease of use, such as the SPICE system (Schultz *et al.* 2007). However these are concept studies, and not generally available, and their



dependence on using servers via the internet make them unsuitable for many places with slow internet connections or no internet connection at all. Consequently, there is a real need and a place for tutorial educational activities focussing on well-tried traditional methodologies such as those discussed in the present report. Bearing in mind the numbers of less resourced languages which do not yet have advanced technological infrastructure, there will continue to be a place for such tutorial educational methods for a long time to come.

## ENDNOTES

1. Science and Technology Education Post-Basic (STEP-B), Federal Government/World Bank Intervention Project: "Towards a Generic Text-To-Speech Applications For African Tone Languages", Grant No.: FME/STEP B/79/3/14 to University of Uyo, Akwa Ibom State, Nigeria, for Moses Ekpenyong and Eno-Abasi Urua.
2. A 'microvoice' is defined as a speech synthesis voice created from a restricted data set without the aim of modelling the entire sound system of a language. Microvoices are typically used in tutorial contexts and in phonetic research contexts for creating synthetic speech for experiments.
3. The well-known MBROLA 'legacy' diphone synthesis front end (<http://tcts.fpms.ac.be/synthesis/>) was chosen, for reasons which will be explained in the text.
4. In the context of a Festschrift contribution for Justus Roux, with whom the first author has had many inspiring discussions in many environments, it is appropriate to combine Justus' interests in speech technology, particularly speech synthesis, African languages, tone and technology infrastructure development in a cooperative advisor-student paper.
5. cf. also the MBROLA web page: <http://tcts.fpms.ac.be/synthesis/>

## REFERENCES

- Akinlabi, A. & M. Liberman. 2000. Tonal Complexes and Tonal Alignment. *North East Linguistics Society (NELS)* 31.
- Bachan, J. 2007. Automatic Close Copy Speech Synthesis. *Speech and Language Technology*, edited by Grażyna Demenko. Poznań: Polish Phonetic Association. Volume 9/10:107-121.
- Bachan, J. 2010. Efficient diphone database creation for MBROLA, a multilingual speech synthesiser. *XII International PhD Workshop (OWD 2010). Conference Archives PTETIS* 28: 303-308.  
[http://www.bachan.speechlabs.pl/files/bachan\\_owd2010.pdf](http://www.bachan.speechlabs.pl/files/bachan_owd2010.pdf) Accessed: 29-03-2011.
- Bachan, J. & D. Gibbon. 2006. Close Copy Speech Synthesis for Speech Perception Testing. *Investigationes Linguisticae* 13:9-24.  
[http://www.staff.amu.edu.pl/~inveling/pdf/Jolanta\\_Bachan\\_Dafydd\\_Gibbon\\_INVE13.pdf](http://www.staff.amu.edu.pl/~inveling/pdf/Jolanta_Bachan_Dafydd_Gibbon_INVE13.pdf) Accessed: 29-03-2011.
- Bloomfield, L. 1933. *Language*. New York: Henry Holt.
- Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glott International* 5(9/10):341-345.
- Boersma, P. & D. Weenink. 2011. Praat: doing phonetics by computer [Computer program]. Version 5.2.21. <http://www.praat.org/> Accessed: 29-03-2011.
- Connell, B. 2002. Downdrift, downstep, and declination. *Proceedings of the Typology of African Prosodic Systems (TAPS) Workshop, Bielefeld, Germany, March 18-20 2001*, edited by U. Gut & D. Gibbon. Bielefeld: Universität Bielefeld.  
<http://www.spectrum.uni-bielefeld.de/TAPS>. Accessed: 2-03-2011.
- Duruibe, U.V. 2010. *A Preliminary Igbo text-to-speech application*. BA thesis. Ibadan: University of Ibadan.
- Dutoit, T. 1997. *An Introduction to Text-To-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers.
- Dutoit, T. & V. Pagel. 1996. Le projet MBROLA: vers un ensemble de synthétiseurs vocaux disponibles gratuitement pour utilisation non-commerciale. *Actes des Journées d'Etudes sur la parole*. Avignon, France.441-444.
- Gibbon, D. 1987. Finite State Processing of Tone languages. *Proceedings of the EACL '87 Conference*. Copenhagen, Denmark.291-297.

- Gibbon, D. 2001. Finite state prosodic analysis of African corpus resources. *Proceedings of the Eurospeech Conference 2001*. Aalborg, Denmark. Volume 1:83-86.
- Gibbon, D. 2009. Prosodic Rank Theory: on the formalisation of prosodic events. *Language, Science and Culture. Essays in Honor of Professor Jerzy Bańczerowski on the occasion of his 70th Birthday*, edited by P. Łobacz, P. Nowak & W. Zabrocki. Poznań: Adam Mickiewicz University Scientific Press.93-126.
- Gibbon, D., R. Moore & R. Winski. (Eds.) 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.
- Gibbon, D., I. Mertins & R. Moore. (Eds.) 2000a. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Dordrecht: Kluwer Academic Publishers.
- Gibbon, D., E.-A. Urua & U. Gut. 2000b. How low is floating low tone in Ibibio? *Proceedings of the 30th Conference on African Languages & Linguistics*. Leiden, Netherlands, 2-30 August 2000.
- Gibbon, D., F. Romani Fernandes & T. Trippel. 2006a. A BLARK extension for temporal annotation mining. *Proceedings of the Language Resources and Evaluation Conference (LREC) 2006*. Genoa, Italy.
- Gibbon, D., E.-A. Urua & M. Ekpenyong. 2006b. Problems and solutions in African tone language Text-To-Speech. *Proceedings of the ISCA Workshop on Multilingual Speech and Language Processing (MULTILING-2006)*, edited J.C. Roux. Stellenbosch, South Africa.
- Gibbon, D. & J. Bachan. 2008. An automatic close copy speech synthesis tool for large-scale speech corpus evaluation. *Proceedings of the Language Resources and Evaluation Conference (LREC) 2008 held in Marrakech, Morocco*, edited by K. Choukri. Paris: ELDA.902-907.
- Gibbon, D., P. K. S. Pandey, D. M. K. Haokip & J. Bachan. 2009. Prosodic issues in synthesising Thadou, a Tibeto-Burman tone language. *Proceedings of the Interspeech Conference 2009*. 6-10 September 2009. Brighton, UK.
- Hirst, D. & A. di Cristo. (Eds.) 1998. *Intonation Systems. A Survey of Twenty Languages*. Cambridge: Cambridge University Press
- Ikekeonwu, C. I. 1993. Intonation and Focus: A Reanalysis of Downdrift and Downstep in Igbo. *Lund University, Dept. of Linguistics Working Papers* 40:95-113.

- Krauwer, S. 2005. ELSNET and ELRA: a common past and a common future. *ELRA Newsletter* 3:2. <http://www.elda.org/article48.html> (14.10.2005 06:27:33). Accessed: 29-03-2011.
- Lieberman, M, Y. & J. B. Pierrehumbert. 1984. Intonational Invariance under Changes in Pitch Range and Length. *Language Sound Structure*, edited by M. Aronoff and R.T. Oehrle. MIT Press.157-233.
- Roux, J., P. Scholtz, D. Klop, C. Povlsen, B. Jongejan & A. Magnúsdóttir. 2010. Incorporating Speech Synthesis in the Development of a Mobile Platform for e-learning. *Proceedings of the Language Resources and Evaluation Conference (LREC) 2010*. Valletta, Malta: European Language Resources Association (ELRA).
- Schultz, T., A.W. Black, S. Badaskar, M. Hornyak & J. Kominek. 2007. SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems. *Proceedings of the Interspeech 2007 Conference*. Antwerp, Belgium, August 2007.

## INDEX

- acoustics, 39, 41, 85
- African languages, 33, 36, 99, 100, 102, 111, 112, 123, 145, 151, 203, 232, 242, 253, 272, 273, 279, 287, 301, 302, 342, 355, 356
- African linguistics, 33, 100, 101
- Afrikaans, 1, 201, 253, 254, 257, 272, 276, 278, 281, 282, 283, 287, 297, 301, 341
- age, 50, 62, 185, 297
- alphabet, 90, 145, 146, 147, 149, 150, 151, 158, 238, 240, 241
- articulatory, 33, 34, 37, 38, 39, 41, 42, 43, 84, 85
- ASR Accuracy, 318
- ASR Correct, 318
- ASR system, 311, 316, 323, 324
- automatic assessment, 309, 312, 318, 324
- automatic speech recognition, 34
- Bantu, 3, 4, 10, 13, 21, 34, 35, 38, 39, 40, 43, 44, 47, 63, 64, 123, 124, 126, 127, 128, 129, 137, 140, 141, 145, 146, 148, 149, 224
- Berlin missionaries, 145, 146
- bilingualism, 10, 163, 164, 165, 166, 167, 168, 169, 172
- bootstrapping, 123, 124, 125, 126, 127, 129, 130, 134, 136, 138, 140, 141
- Botswana, 39, 62, 128, 220, 230, 231
- BSAE, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 299
- burst amplitude, 21, 22, 25, 29
- Burst Amplitude*, 24
- business drivers, 293, 295, 296, 302
- Chadic, 99, 100, 111
- Civili, 83, 84, 85, 86, 87, 88, 89, 90, 92
- closure duration, 22
- collocation*, 179, 180, 183, 184, 185, 230, 328
- communication, 35, 85, 163, 169, 170, 172, 177, 178, 199, 200, 201, 202, 203, 204, 207, 209, 212, 291, 292, 311, 328, 331, 334
- Computer applications, 211
- concordance, 217, 219, 220, 221, 222, 223, 225, 232
- consonants, 4, 6, 39, 42, 43, 61, 62, 64, 65, 66, 69, 70, 71, 73, 76, 77, 79, 80, 89, 154, 155, 156, 159, 160
  - syllabic consonants, 61, 62, 64, 65, 66, 79
- corpus, 2, 4, 5, 10, 12, 23, 24, 30, 37, 38, 83, 86, 87, 88, 90, 102, 110, 128, 140, 179, 201, 202, 206, 217, 218, 219, 222, 223, 229, 230, 232, 238, 239, 243, 246, 272, 274, 275, 276, 277, 278, 279, 288, 316, 328, 343, 344, 345, 350, 351, 352
- culture, 99, 100, 108, 173, 204, 238, 255, 299, 302
- Danish, 328, 330, 331, 332, 334, 336
- data, 2, 3, 4, 7, 9, 10, 11, 12, 13, 23, 24, 37, 38, 40, 43, 44, 58, 62, 66, 75, 79, 83, 84, 85, 86, 87, 89, 92, 138, 159, 165, 177, 181, 200, 205, 206, 209, 217, 221, 222, 239, 240, 251, 255, 273, 274, 275, 276, 277, 278, 279, 280, 281, 286, 287, 288, 292, 293, 299, 310, 311, 316, 317, 328, 329, 330, 333, 336, 339, 340, 343, 349, 352, 356
  - data selection, 273, 275, 276, 277, 278, 279, 280, 281, 286, 288
  - frequency-based, 273, 281
  - phonetic data, 35
  - speech data, 33, 85, 86, 89, 340
- diacritics, 146, 147, 151, 152, 153, 157, 158, 159
- dictionary basis, 237, 239
- dictionary database, 217
- double-layered sublemmata, 265, 266, 267
- duration, 1, 21, 22, 29, 40, 43, 48, 49, 83, 85, 86, 87, 88, 89, 92, 240, 317, 343, 344
- education, 34, 111, 165, 166, 168, 169, 172, 173, 202, 203, 205, 206, 207, 309, 310, 312, 313, 339, 355
- emerging market, 292, 302

- encyclopaedic dictionary, 237, 239, 241, 243, 246, 247
- English, 1, 3, 4, 8, 10, 11, 12, 17, 35, 38, 50, 87, 149, 150, 163, 164, 167, 168, 173, 177, 178, 179, 181, 182, 184, 187, 188, 201, 207, 216, 271, 276, 277, 278, 279, 281, 282, 283, 284, 285, 286, 287, 295, 297, 300, 302, 303, 310, 311, 312, 313, 314, 316, 320, 331
- explosive, 21, 22, 24, 25, 26, 27, 28, 29
- first language, 4, 57, 169, 171, 172
- French, 1, 10, 11, 35, 87, 163, 165, 166, 167, 168, 169, 171, 172, 173, 177, 181, 182, 184, 188, 189, 207, 238, 239, 241, 242, 328
- fundamental frequency, 11, 48, 49, 56, 89, 347
- Gabon, 163, 165, 166, 167, 168, 170, 171, 172, 181, 186, 237, 238, 239, 242, 243
- Gender, 298
- German, 1, 13, 100, 101, 112, 128, 146, 149, 150, 165, 167, 168, 172, 178, 188, 278, 304, 328, 329
- grapheme, 156
- harmonisation, 138, 203, 204
- human language technologies, 38, 44, 200, 206
- idiom, 146, 177, 178, 184, 188, 225
- implosive, 21, 22, 24, 25, 26, 27, 28, 29, 30
- information science, 327, 336
- information tool, 331, 333
- intensity, 1, 11, 24, 25, 29, 39, 42, 48, 49, 140, 217
- internet lexicography, 329
- isiZulu, 272, 276, 278, 281, 283, 284, 286, 287, 288, 297, 303
- knowledge transfer, 200
- Lamang*, 99, 100, 101, 102, 103, 104, 105, 106, 108, 110, 111, 112, 114, 116
- language development, 34, 200
- language transfer, 149
- Latin, 145, 151, 158, 181, 257
- learning rates, 277, 279, 286, 287
- lemma parts, 258, 262, 263, 264
- lemmatisation procedures, 256, 258, 265
- lexicography, 128, 165, 200, 202, 203, 205, 208, 210, 211, 225, 232, 239, 240, 251, 265, 267, 327, 328, 329, 336
- linguistic ability, 311, 318, 321
- linguistic environment, 168, 169, 170, 173
- macrostructure, 251, 252, 253, 254, 255, 256, 259, 261, 262, 265, 267
- minimal pairs, 1, 23, 27, 37, 86, 90, 353
- modernisation, 201
- morphological analysis, 123, 124, 129, 141
- multilingualism, 163, 165, 173, 210, 296
- Ndebele, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 137, 138, 140, 141
- nesting, 252, 254, 257, 260, 261, 265
- Ngoni, 124, 126, 127, 128, 129, 134, 135, 136, 137, 138, 139, 140, 141
- Nguni, 2, 22, 23, 24, 26, 27, 29, 123, 124, 125, 126, 127, 128, 133, 140, 141, 150, 156, 287, 288
- niching, 252, 265
- Nigerian languages, 339
- non-grouped lemmata, 265
- Northern Sotho, 47, 145, 146, 147, 148, 149, 150, 152, 154, 156, 158, 159, 161
- oral proficiency, 309, 310, 312, 313, 314, 320, 321, 322, 323, 324
- oral proficiency assessment, 309, 324
- orature-grammar interface, 111
- orthography, 6, 7, 8, 9, 35, 36, 90, 108, 136, 146, 147, 148, 149, 150, 151, 152, 153, 155, 156, 158, 159, 161, 200, 202, 203, 240, 241, 257
- partial lemmata, 257, 258, 260, 261, 262, 265, 266, 267
- Persona design, 297
- Phonetics, 42, 58, 84
- phonetic representation, 92, 150, 350
- Phonetics-Phonology Interface Debate*, 83, 84
- Phonology, 38, 58, 61, 84
- Distinctive Feature Phonology, 61
- phonological processes, 61, 66, 67, 73, 79, 80
- phonological theory, 38, 44, 92

- proverbs, 100, 102, 103, 110, 178, 180, 181, 184, 185, 186, 189, 216, 220, 228, 232, 254
- pulmonic, 21
- resources
  - language resources, 123, 206, 207, 208, 209, 301, 302
- resource-scarce languages, 123, 125, 140, 273, 275, 279, 287, 288
- rhymes, 99, 100, 102, 103, 106, 108, 111, 181
- second language, 15, 164, 167, 168, 169, 170, 171
- Sepedi, 145, 146, 201, 272, 279, 281, 285, 286, 287, 288
- Sesotho, 47, 57, 61, 62, 63, 64, 65, 67, 72, 73, 74, 75, 76, 77, 79, 201, 288, 297
- Sesotho sa Leboa, 145
- Setswana, 61, 62, 63, 64, 65, 67, 72, 73, 74, 75, 79, 216, 217, 218, 220, 222, 224, 228, 229, 230, 232, 276, 288
- South Africa, 145, 291, 292, 294, 303
- Southern Sotho, 47, 48, 50, 81, 146, 148
- speech, 4, 5, 10, 13, 24, 33, 34, 36, 37, 38, 39, 40, 41, 42, 43, 49, 51, 54, 58, 75, 83, 85, 86, 89, 90, 91, 103, 123, 124, 147, 148, 149, 179, 189, 201, 206, 209, 238, 292, 293, 294, 298, 299, 300, 301, 302, 303, 309, 310, 312, 316, 317, 320, 322, 323, 324, 327, 339, 340, 341, 342, 343, 345, 347, 350, 355, 356
- speech synthesis, 34, 90, 327, 339, 340, 341, 342, 355, 356
- standard language, 145, 201, 203
- standardisation, 36, 52, 146, 147, 148, 152, 199, 200, 201, 202, 203, 204, 205, 207, 208, 209, 210, 211, 212, 247
- statistical machine translation, 271, 272, 273, 274, 275, 276, 278
- stress, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 36, 37, 88, 347
- sublemmata, 253, 257, 258, 259, 261, 262, 263, 264, 265, 266, 267
- Swati, 22, 23, 24, 25, 26, 27, 28, 29, 124, 125, 129, 140, 141
- syllabification, 61, 66, 69, 72, 73, 74, 75, 76, 77, 79, 80
- syllable, 1, 3, 5, 6, 7, 8, 9, 11, 12, 13, 37, 47, 62, 63, 64, 72, 76, 77, 88, 353
- syllable weight, 1, 3
- terminography, 200, 202, 203, 204, 205, 207, 208, 211
- terminology, 38, 42, 165, 167, 180, 199, 200, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 332
- tone, 1, 10, 11, 13, 36, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 157, 158, 293, 301, 302, 341, 342, 346, 347, 348, 353, 355, 356
- tone languages, 11
- touch-tone, 293, 300, 302
- vowel duration, 84, 85, 86, 89, 90, 92
- vowel formant, 48, 49
- vowel quality, 5, 8, 11, 13, 47, 48, 49, 51, 53, 54, 55, 56, 57
- Xhosa, 22, 23, 24, 25, 26, 27, 28, 29, 124, 125, 127, 129, 140, 141
- Yilumbu*, 177, 180, 181, 185, 186, 187, 188, 189, 237, 238, 239, 240, 241, 243, 246, 247
- Zambia, 39, 126, 128
- Zulu, 21, 22, 23, 24, 25, 27, 29, 30, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 147

*This book provides a broad overview of current work on South African languages, language resources and language technologies. While it provides a fairly comprehensive overview, it also ties together the most recent knowledge state here, and is therefore truly innovative ... The book is therefore informed by current international trends in the respective fields of science, and feeds back into them ... There is absolutely no doubt that the book has an academic peer audience and is directed at specialists in the field.*

**Prof. Axel Fleisch**, University of Helsinki, Finland

*This Festschrift, containing 20 stimulating articles by colleagues and students, is a fitting tribute to his 65th birthday. The book covers a wide range of subjects dealing with phonetics and phonology, language description and resources, lexicography and terminology, and language technology research and development. Each article presents the results of experiments completed or the exposition of viewpoints held in connection with these subject fields. However, the presentations do not just end here. They also contain pointers to further problems for research and discussion. These articles are therefore not complete in themselves, but open-ended, leaving not only the possibility for reconsidering the research and discussions presented, but also, because the information and viewpoints given are original and innovative, for stimulating thought and evoking reaction. These articles are therefore meant for specialists in the relevant fields who can fully appreciate the experiments and arguments and respond to them in an informed manner.*

**Dr Johan du Plessis**, Bureau of the WAT, Stellenbosch, South Africa

---

**Dr H. Steve NDINGA-KOUMBA-BINZA** is Research Fellow at the Centre for Text Technology, North-West University in Potchefstroom, South Africa, and Assistant Professor of phonetics and phonology at Omar Bongo University in Libreville, Gabon. His research interests include African vowel systems and syllable structures for use in text and speech technology development and writing and spelling systems. He has authored and co-authored numerous publications, including the books *Civili, langue des Baloango: Esquisse historique et linguistique* (Lincom Europa 2010; co-author: PA Mavoungou) and *A phonetic and phonological account of the Civili vowel duration* (Cambridge Scholars Press 2012). He received his Doctor Litteratum (PhD) degree under Justus C. Roux in 2008 at Stellenbosch University, South Africa.

**Prof. Sonja E. BOSCH** is Chair of the Department of African Languages at the University of South Africa in Pretoria. She teaches Zulu acquisition courses and undergraduate as well as postgraduate courses in morphology and syntax. She is project leader of a research group dealing with computational morphological analysis of African languages, as well as coordinator of the African Wordnet project. As researcher, she has organised several workshops nationally and internationally and is an NRF-rated researcher. Her publications include more than 50 peer-reviewed journal articles, conference proceedings and book chapters, as well as three books published nationally and internationally. She has been a friend, a colleague and co-project leader to Justus C. Roux for many years.



[www.sun-e-shop.co.za](http://www.sun-e-shop.co.za)